Isolation and Characterisation of Four Cathepsin B-like Cysteine Protease Genes from the Free Living Nematode Caenorhabditis elegans

Christopher Geoffrey Carson Larminie

Ph.D.
University of Glasgow
Wellcome Unit of Molecular Parasitology
December 1995
Contents

Abstract i
Acknowledgements ii
Abbreviations iii

Chapter 1
General Introduction

1.1. The Phylum Nematoda 1
1.2. Caenorhabditis elegans biology 1
1.2.1. C.elegans development 2
1.2.2. C.elegans anatomy 4
1.2.3. C.elegans genetics 6
1.3. C.elegans as an experimental system 6
1.4. C.elegans as a model organism 8
1.5. Cathepsin B 10
1.5.1. Papain 10
1.5.2. Isolation and biochemical characterisation of cathepsin B 12
1.5.3. Processing of cathepsin B 15
1.6. Tissue distribution of cathepsin B 17
1.7. The biological roles of cathepsin B 17
1.7.1. A lysosomal role for cathepsin B in protein turnover and degradation 17
1.7.2. A role for cathepsin B in the processing of protein precursors 19
1.7.3. A role for cathepsin B in antigen processing 20
1.7.4. A role for cathepsin B in bone remodelling 21
1.7.5. A pathological role for cathepsin B in tumour cell invasion 23
1.8. Cathepsin B and parasitism 25
1.9. Cathepsin B-like multigene families 30
1.10. Cathepsin B-like enzymes in C.elegans 31
1.11. Project Aims and Objectives 32

Figures for Chapter 1

Chapter 2
Materials and Methods

2.1. Chemical abbreviations 36
2.2. Commonly used stocks, solutions and media
2.3. *C. elegans*, yeast and bacterial strains
2.4. Vectors and clones
2.4.1. Plasmid vectors
2.4.2. Phage
2.4.3. Clones
2.5. Culture maintenance
2.5.1. *C. elegans* culture
2.5.2. Yeast culture
2.5.3. Bacterial culture
2.5.3.1. Bacterial culture on plates
2.5.3.2. Bacterial liquid culture
2.6. DNA preparation
2.6.1. Plasmid DNA mini preparation
2.6.2. ssDNA template preparation for sequencing
2.6.3. Plasmid DNA medium and large scale preparations
2.6.4. Large scale cosmId DNA preparations
2.6.5. Lambda phage DNA preparations
2.6.6. Yeast genomic DNA preparation
2.7. RNA preparation
2.7.1. Isolation of mixed stage embryos
2.7.2. RNA extraction
2.8. Standard molecular biology techniques
2.8.1. Restriction endonuclease digests
2.8.2. Agarose Gel electrophoresis of RNA and DNA
2.8.2.1. Agarose gels for electrophoresis of DNA samples
2.8.2.2. Agarose gels for electrophoresis of RNA samples
2.8.3. Gel purification
2.8.4. Transformation of *E. coli*
2.8.5. Ligation reactions
2.8.6. Southern and Northern blots
2.8.7. Probe preparation
2.8.7.1. Random priming
2.8.7.2. 5' end labelling
2.8.7.3. Riboplates
2.8.7.4. Probe purification
2.8.8. Hybridisation
2.8.8.1. Southern hybridisations
2.8.8.2. Northern hybridisations
2.8.8.3. Southern hybridisation washes
2.8.8.4. Northern hybridisation washes
2.8.8.5. Autoradiography
2.8.8.6. Colony hybridisation
2.9. Screening a lambda EMBLA C.elegans
   Bristol N2 genomic DNA library
2.9.1. Determining the titre of the lambda library
2.9.2. Plating and screening of the lambda library
2.10. Excision of cDNA clones from lambda SHLX2
2.11. Nucleotide sequencing
2.11.1. Preparation of custom made primers
2.11.2. Deprotection of synthesised oligonucleotides
2.11.3. Nucleotide sequence of the primers used for sequencing
2.11.3.1. Primers for sequencing the 5' rapid amplification of
cDNA ends (5' RACE) products
2.11.3.2. Primers for sequencing the cDNA and genomic clones
   of cpr-3, cpr-4, cpr-5 and cpr-6
2.11.3.3. Primers for sequencing the cpr-3::lacZ, cpr-4::lacZ,
cpr-5::lacZ and cpr-6::lacZ constructs
2.12. 5' Rapid amplification of cDNA ends (5' RACE)
2.12.1. Reaction conditions of the nested PCR amplification steps
2.12.1.1. First round of nested PCR
2.12.1.2. Second round of nested PCR
2.13. Semi-quantitative reverse transcription
   polymerase chain reaction (s-q rtPCR)
2.13.1. PCR primers used for s-q rtPCR
2.13.2. The PCR reaction conditions used for s-q rtPCR
2.13.3. PCR amplification of sequences required to
   probe the s-q rtPCR Southern blots
2.14. Colony PCR
2.15. High fidelity PCR
2.16. PCR amplification of the 5' ends of the cloned cDNAs
   of cpr-3, cpr-4, cpr-5 and cpr-6
2.17. Generation of transgenic C.elegans lines and
   staining for β-Galactosidase activity
2.17.1. Preparation of materials and worms for microinjection
2.17.2. Microinjection procedure
Chapter 3
Cloning Four Cathepsin B-Like Genes From *C. elegans*

3.1. Introduction

3.1.1. The *C. elegans* physical map

3.1.2. The *C. elegans* genome sequencing project

3.1.3. Additional uses of the *C. elegans* physical map

3.2. Results

3.2.1. Obtaining four cDNA clones encoding proteins with homology to cathepsin B

3.2.2. Confirming the physical map positions of *cpr-3, cpr-4, cpr-5* and *cpr-6*

3.2.3. Analysis of the Southern blots of *C. elegans* genomic DNA and yeast DNA

3.2.3.1. The hybridisation pattern for *cpr-3*

3.2.3.2. The hybridisation pattern for *cpr-4*

3.2.3.3. The hybridisation pattern for *cpr-5*

3.2.3.4. The hybridisation pattern for *cpr-6*

3.2.4. The copy number of *cpr-3, cpr-4, cpr-5* and *cpr-6*

3.2.5. *cpr-3, cpr-4, cpr-5* and *cpr-6* do not cross hybridise significantly

3.2.6. Selecting cosmid clones for *cpr-3, cpr-4, cpr-5* and *cpr-6*

3.2.7. Isolating genomic clones containing *cpr-3, cpr-4, cpr-5* and *cpr-6*

3.2.8. Isolation of subclones from the positive cosmid and lambda clones

3.2.8.1. Obtaining suitable fragments for subcloning

3.2.8.2. Subcloning fragments containing *cpr-4, cpr-5* and *cpr-6*

3.2.8.3. Isolating the *cpr-3* gene from lambda clones

3.2.8.4. Obtaining subclones containing *cpr-3* from CL#034

3.2.9. Restriction endonuclease mapping of the genomic clones

3.2.9.1. The position, orientation, maximum size and copy number of *cpr-3, cpr-4, cpr-5* and *cpr-6* within their subclones
3.2.9.2. The nucleotide sequences of cpr-3, cpr-4, cpr-5 and cpr-6 support the restriction mapping data
3.2.10. Subcloning the 5′ flank of cpr-3 from CL#034
3.2.11. The generation of subclones from ZK1055/6.8/KS-, W02B2/11/KS- and C25B8/10.7/KS-
3.2.12. Determining whether the subclones are contigous with the C. elegans genome
3.2.13. The physical map positions of cpr-3, cpr-4, cpr-5 and cpr-6
3.3. Discussion
Figures for Chapter 3

Chapter 4
Characterisation of cpr-3, cpr-4, cpr-5 and cpr-6

4.1. Introduction
4.2. Results
4.2.1. Confirming the identity of cpr-3, cpr-4, cpr-5 and cpr-6
4.2.2. Sequencing the cDNA clones of cpr-3, cpr-4, cpr-5 and cpr-6
4.2.3. Obtaining the full-length mature transcripts of cpr-3, cpr-4, cpr-5 and cpr-6
4.2.4. Analysis of the full-length mature transcripts of cpr-3, cpr-4, cpr-5 and cpr-6
4.2.5. The gene architectures of cpr-3, cpr-4, cpr-5 and cpr-6
4.2.6. Putative regulatory elements in the 5′ flanks of cpr-3, cpr-4, cpr-5 and cpr-6
4.2.7. cpr-3, cpr-4, cpr-5 and cpr-6 encode proteins with homology to cathepsin B
4.2.8. Phylogenetic analysis
4.2.8.1. The predicted proteins encoded by cpr-3, cpr-4, cpr-5 and cpr-6 cluster with known cathepsin B enzymes
4.2.8.2. Phylogenetic comparison of the proteins encoded by cpr-3, cpr-4, cpr-5 and cpr-6 with cathepsin B-like sequences in the database
4.2.8.3. Gene architecture as a gauge of evolutionary divergence
4.3. Discussion
4.3.1. A cathepsin B-like multigene family in C. elegans
4.3.2. The Implications of isolating a cathepsin B-like multigene family from C. elegans
4.3.2.1. *C. elegans, H. contortus* and *O. ostertagi* may possess more as yet unidentified cathepsin B-like genes

4.3.2.2. The cathepsin B-like multigene families of free living and parasitic nematode species may have arisen from different gene duplication events

4.3.2.3. Cathepsin B-like multigene families may be present in other metazoans

4.3.3. The cathepsin B-like genes of *C. elegans* may encode enzymes with different biological functions

4.3.3.1. *cpr-1, cpr-3, cpr-4, cpr-5* and *cpr-6* may encode enzymes with different specificities and activities

4.3.3.2. The enzymes encoded by *cpr-3, cpr-4, cpr-5* and *cpr-6* may be subjected to different types of post-translational modification

4.3.4. Regulation of expression of *cpr-3, cpr-4, cpr-5* and *cpr-6*

4.3.4.1. Post-transcriptional regulation

4.3.4.2. Expression of *cpr-1, cpr-3, cpr-4, cpr-5* and *cpr-6* may be temporally and/or spatially restricted

Tables for Chapter 4

Figures for Chapter 4

Chapter 5

The Temporal and Spatial Expression Patterns of *cpr-3, cpr-4, cpr-5* and *cpr-6*

5.1. Introduction

5.2. Analysing the temporal expression patterns of *cpr-3, cpr-4, cpr-5* and *cpr-6*

5.2.1. Experimental approaches used

5.2.2. Results of Northern blot analysis

5.2.3. Results from sq-rtPCR analysis

5.2.3.1. Initial analysis of the temporal expression patterns of *cpr-3, cpr-4, cpr-5* and *cpr-6* using sq-rtPCR

5.2.3.2. Adaptations made to the initial s-q rtPCR protocol

5.2.3.3. Summary of the modified s-q rtPCR protocol used

5.2.3.4. The modified s-q rtPCR approach generates reproducible results

5.2.3.5. The temporal expression patterns of *cpr-3, cpr-4, cpr-5* and *cpr-6* obtained from the modified s-q rtPCR approach
5.2.4. Comparison of the results obtained from Northern and s-q rtPCR analysis

5.3. Analysing the spatial expression patterns of cpr-3, cpr-4, cpr-5 and cpr-6

5.3.1. Overview of the lacZ fusion transgene approach

5.3.2. Advantages of using lacZ transgene fusions

5.3.3. Problems associated with the use of lacZ transgene fusions

5.3.4. lacZ fusion vectors

5.3.5. Generation of the lacZ gene fusions

5.3.6. Vector construction

5.3.6.1. Screening the PCR amplified linkers for mutations

5.3.6.2. Preparation of the lacZ reporter constructs

5.3.6.3. Checking the integrity of the lacZ reporter constructs

5.3.6.4. Final analysis of the lacZ reporter constructs prior to injection

5.3.7. Generating C. elegans lines transgenic for the reporter constructs and staining for ß-Galactosidase activity

5.3.8. Gut-specific expression of cpr-3, cpr-4, cpr-5 and cpr-6

5.3.8.1. Variable expression of the lacZ fusion transgenes in the intestine

5.3.8.2. Variability in the number and intensity of staining nuclei may be due to the different lacZ fusion vectors used

5.3.8.3. cpr-3, cpr-4, cpr-5 and cpr-6 may be expressed in different regions of the intestine

5.3.9. The temporal expression patterns indicated from the lacZ fusion analysis correlate with those indicated by s-q rtPCR

5.4. Discussion

5.4.1. Overview of results obtained

5.4.2. Efficacy of the results obtained using lacZ reporter gene fusions

5.4.3. Potential biological roles for cpr-3, cpr-4, cpr-5 and cpr-6

5.4.4. Biological interaction

Tables for Chapter 5

Figures for Chapter 5
Chapter 6
Concluding Discussion

6.1. Summary of results

6.2. Future work

6.2.1. Determining the distribution of the cathepsin B-like enzymes

6.2.2. Biochemical characterisation of the cathepsin B-like enzymes

6.2.3. The effect of starving on expression of \textit{cpr-3}, \textit{cpr-4}, \textit{cpr-5} and \textit{cpr-6}

6.2.4. Obtaining \textit{C.elegans} mutant strains for \textit{cpr-3}, \textit{cpr-4}, \textit{cpr-5} and \textit{cpr-6}

References

Appendix A
Abstract

The cathepsin B cysteine protease enzyme performs a role in protein turnover and degradation within the lysosomes of vertebrates. This proteolytic enzyme is also thought to perform related roles in the processing of antigens and protein precursors, as well as a role in bone resorption. A pathological role for cathepsin B in tumour cell invasion has also been suggested. Enzymes with cathepsin B-like activities are thought to be excreted and/or secreted by a variety of parasitic nematode and trematode species. To date, multigene families with the potential to encode cathepsin B-like enzymes have only been reported in parasitic nematode and trematode species, suggesting that these enzymes may be important for parasitism by these species. The work presented in this thesis demonstrates that the free living nematode species, Caenorhabditis elegans, also possesses a cathepsin B-like multigene family indicating that such multigene families are not unique to parasitic nematode and trematode species. Four genes with homology to vertebrate cathepsin B were isolated from the genome of C. elegans and were named cpr-3, cpr-4, cpr-5 and cpr-6. Phylogenetic analysis clusters the proteins encoded by these four genes with known cathepsin B enzymes and away from other, related enzymes such as cathepsins H and L. Since the four genes possess distinct genomic architectures, they appear to have arisen from ancient gene duplication events. This is supported by phylogenetic analysis which clusters the predicted proteins encoded by these four genes and cpr-1, a previously isolated C. elegans cathepsin B-like gene (Ray and McKerrow, 1991), into three groups which are almost as diverged from one another as each is to the vertebrate cathepsin B enzymes. The expression of cpr-3, cpr-4, cpr-5 and cpr-6 as lacZ reporter transgene fusions in transgenic worms suggests that these four genes are all exclusively expressed in the intestinal cells of C. elegans. Analysis of the temporal expression patterns of these four genes using semi-quantitative reverse transcription polymerase chain reaction indicates that the four genes exhibit distinct but overlapping temporal patterns of expression during C. elegans development. These results suggest a differential requirement for cathepsin B-like enzymes, or combinations of enzymes, within the intestine during C. elegans development.
Acknowledgements

I would like to thank Dr. Iain Johnstone, my supervisor, for his encouragement and support over the last three years. I would also like to thank the other people involved in C.elegans research at Glasgow University and in particular Dr. John 'Boy' Gilleard. I am also grateful to the people of the Wellcome Unit of Molecular Parasitology with whom I have enjoyed working. I would like to acknowledge the financial support of the Scottish Office Agriculture and Fisheries Department. I am especially grateful to A.J.M. for all her support over the last three years, particularly during the writing of this thesis.

Declaration

This thesis and the results presented in it are entirely my own work, except where indicated.
Abbreviations

bp base pair
BSA bovine serum albumen
cDNA complementary deoxyribonucleic acid
cm centimetre
cpm counts per minute
dna deoxyribonucleic acid
E/S excretory/secretory
EST expressed sequence tag
IVR in vitro released
kb kilobase
kcal kilocalories
kDa kiloDalton
l litre
lbs pounds
M Molar
Mb megabase
MCS multiple cloning site
min minute
ml millilitre
mM Millimolar
mm millimetre
μm micrometre
μl microlitre
μg microgram
ng nanogram
OD optical density
PCR polymerase chain reaction
pfu plaque forming unit
pg picogram
RACE rapid amplification of cDNA ends
rDNA ribosomal deoxyribonucleic acid
RNA ribonucleic acid
rpm revolutions per minute
RT-PCR reverse transcription polymerase chain reaction
s-q rtPCR semi-quantitative reverse transcription polymerase chain Reaction
ssDNA single stranded Deoxyribonucleic acid
UTR untranslated region
UV ultraviolet
V volt
YAC yeast artificial chromosome
X-gal 5-bromo-4-chloro-3-indolyl-β-D-galactopyranoside
Chapter 1
Chapter 1

General Introduction

1.1. The Phylum Nematoda

The phylum Nematoda is large, containing 10,000 described nematode species, and its members are among some of the most widespread and numerous of all multicellular animals. Members of this phylum extend from the polar regions to the tropics and inhabit all types of marine and terrestrial environments including; high altitudes, hot springs, deserts and oceans. Parasitic nematodes are equally diverse, displaying all the different degrees of parasitism and attacking virtually all groups of plants and animals. Despite the diverse environments and very different lifestyles of different nematode species, all are morphologically and anatomically very similar. All nematode species have a long slender body and a circular cross section (hence the name ‘roundworms’). The general body plan comprises two concentric tubes separated by the pseudocoelomic space. The outer tube consists of the cuticle, the hypodermis, muscle and nerve cells while the inner tube consists of the intestine and its lumen. The gonads of adult worms are contained in the pseudocoelomic space. For a more in-depth overview of the characteristics of the phylum Nematoda, see Barnes (Barnes, 1980).

1.2. Caenorhabditis elegans biology

The free-living soil nematode, Caenorhabditis elegans, has been the subject of extensive studies and is the most well characterised of all nematode species. It is a filter feeder whose diet consists primarily of bacteria. C.elegans can be easily maintained in the laboratory on agar plates or in liquid culture using Escherichia coli as a food source (Brenner, 1974) and the life cycle takes only approximately 3½ days to complete under optimal conditions. The two sexes, hermaphrodite and male, are a similar size, approximately 1 mm in length and 70 μ in diameter. However, the two sexes can be distinguished by virtue of different anatomical features (see Section 1.2.2). C.elegans propagates primarily by hermaphroditic self-fertilisation since hermaphrodites cannot
fertilise one another and males, which are capable of fertilising hermaphrodites, occur in the population only at a low frequency (approximately 1/700).

1.2.1. *C. elegans* development

The life cycle of *C. elegans*, like all other nematode species, involves four larval stages, named L₁ - L₄. Eggs are fertilised internally and early embryogenesis occurs within the parent. The later stages of embryogenesis occur after the eggs have been laid. Embryogenesis is rapid, taking only approximately 14 hours to complete in optimal conditions. After hatch, the morphology of the resulting 1st stage larvae (L₁) is superficially quite similar to the adult. Three further larval stages occur (L₂, L₃ and L₄) prior to generation of the adult approximately three days after hatch. The end of each larval stage is punctuated by a moult. At each moult, the old cuticle is shed to reveal a newly synthesised cuticle beneath. During shedding of the old cuticle, pharyngeal pumping stops for a brief period known as lethargus. Under conditions of overcrowding or starvation, the L₂ larvae can enter an alternative developmental pathway to generate dauer larvae which are long-lived and resistant to unfavourable conditions (Riddle, 1988). In more favourable conditions, these dauer larvae moult to L₄ stage larvae and resume normal development.

The body wall of *C. elegans* and the shell of eggs and developing embryos are transparent. This feature, in conjunction with Nomarski microscopy (a high resolution, non-destructive light microscopy technique), has made it possible to describe the complete embryonic (Sulston *et al.*, 1983) and post-embryonic (Sulston and Horvitz, 1977) cell lineages of *C. elegans*. These studies followed individual nuclei in live animals and compared the resulting observations to the cellular anatomy of the worm, determined at both the light and electron microscope levels. These studies have revealed that both the embryonic and post-embryonic cell lineages are highly invariant, with each lineage rigidly determined and giving rise to a fixed number of cells of strictly specified fates. Although the relationship between cell ancestry and cell fate is fixed, there is little correlation between them. Thus, the six founder cells, AB, MS, E, C, D and P₄, generated by the first four cleavages, do not correspond exactly to specific germ layers. However, three founder cells do give rise to ‘pure’ clones; E gives rise exclusively to the
intestine, D produces body wall muscle (but not all body wall muscle) and P₄ gives rise to the germ-line.

As a result of the cell lineage studies described above, the entire period of *C.elegans* development is very well characterised. Embryogenesis can be divided into two stages of approximately the same length; first, the cell proliferation and organogenesis stage and second, the morphogenesis stage. During the first stage, cell divisions, cell movements and some cell deaths take place. By the end of this phase the gastrulated embryo is spheroid with a fixed number of cells (approximately 550) whose fates are rigidly determined. During the second phase, cell proliferation stops almost completely and morphogenesis occurs. During this stage, the embryo elongates more than threefold, neural processes grow out and interconnect, the embryo begins to move actively and cuticle synthesis occurs.

The L₁ stage larvae comprise approximately 550 somatic nuclei in both hermaphrodites and males. Both sexes possess a gonadal primordium at this larval stage which comprises four cells. The number of somatic nuclei increases to 959 in the mature adult hermaphroditic and 1031 in the mature adult male. These additional nuclei are generated both by division of non-gonadal somatic blast cells and by proliferation of the gonadal primordium during post-embryonic development.

The non-gonadal somatic blast cells make up approximately 10% of the non-gonadal nuclei of both hermaphrodite and male L₁ stage larvae (Sulston and Horvitz, 1977). As a result of division of these blast cells, the number of non-gonadal somatic nuclei increases to about 810 in the mature adult hermaphrodite and 970 in the male. In both sexes, the daughter cells arising from these divisions contribute to the hypodermis, nervous system, intestine and musculature. Most of the additional male-specific cells are located in the specialised structures of the male tail which are required for copulation.

The remaining somatic nuclei of the mature adult hermaphrodite and mature adult male are generated by proliferation of the gonad. The hermaphrodite and male gonad develops during the larval stages, from a gonadal primordium comprising four cells in the L₁ stage. Of these four morphologically similar cells, two contribute to the somatic gonad and two to the germ-line tissue. The somatic gonad of both sexes develops according to an invariant pattern of cell divisions (Kimble and Ward, 1988), resulting in approximately 140 nuclei contributing to the hermaphrodite gonad and 60 nuclei to the
male gonad. Though gonadogenesis occurs throughout larval development, the structures of the gonad only begin to be easily visible around the late L₃ stage. Sexual maturation continues to the end of the L₄ stage, culminating in the opening of the hermaphrodite vulva and morphogenesis of the male tail at the final moult.

1.2.2. *C. elegans* anatomy

The complete anatomy of *C. elegans* has been determined at the cellular level mostly using reconstructions of electron micrographs of serial sections. This information, in conjunction with the invariance of cell number and cell fate (discussed above), has allowed every cell in the worm to be identified and given a unique label. The anatomy of *C. elegans* is typical of all nematode species and essentially consists of two concentric tubes separated by pseudocoelomic space. The outer tube is made up of a cuticle, hypodermis, neurones and muscle which all surround the pseudocoelom. The inner tube consist of the intestine and its lumen. The gonads of both sexes are contained within the pseudocoelomic space (Figure 1.1A).

The cuticle is made up of collagen, organised into three main layers, and is synthesised by an underlying external epithelium called the hypodermis. Accordingly, the hypodermis extends over the entire surface of the worm, including parts of the lumen of the pharynx and anus. Two elevated longitudinal ridges, called alae, mark the lateral surfaces of the adult cuticle. These alae are synthesised by seam cells, a subset of hypodermal cells which underlie the alae. On solid surfaces, *C. elegans* crawls on one side with the alae contacting the surface of the medium. Movement itself is attained using four strips of striated body wall muscle running along the length of the animal (Figure 1.1A).

The alimentary canal comprises the pharynx, the intestine and the rectum. The pharynx functions to ingest, concentrate and process food prior to pumping it into the intestine. Accordingly, the pharynx is made up of muscles, epithelial cells and nerves. The lumen of the pharynx is lined with cuticle and this cuticle is continuous with the body cuticle. The pharyngeal cuticle possesses several specialised structures the most obvious of which are the knob like structures in the terminal bulb (Figure 1.1B), which grind up food. The pharynx is connected to the intestine via the pharyngeal-intestinal
valve. The intestine is comprised of two rows of eight cells (int2-int9) and an anterior ring of four cells (int1). The four cells comprising int1 represent the point of attachment of the intestine to the cells of the pharyngeal-intestinal valve. The intestine is attached to the rectum via the intestinal-rectal valve, a very similar structure to the pharyngeal-intestinal valve. The rectum itself is made up of three pairs of endothelial cells. The rectum is associated with three sets of muscles required for excretion which are coupled by gap junctions and innervated by a single neuron.

The hermaphrodite and male gonads have very different structures (Figure 1.1B). The hermaphrodite gonad has two lobes, one extending anteriorly and the other posteriorly from the centre of the worm. Each lobe is U-shaped and comprises an ovary (at the distal end), an oviduct and a spermatheca (at the proximal end). At the distal end, the ovaries are syncytial and contain germ-line nuclei. As the nuclei move proximally, they progress through the stages of meiosis and are in diakinesis by the time they reach the oviduct. By this stage, each of the germ-line nuclei are enclosed by membranes to form large oocytes. The oviduct terminates at the spermatheca, containing amoeboid sperm. The two spermathecae (one from each lobe) are connected by a common uterus which contains fertilised eggs. The uterus opens to the outside via a vulva protruding from the ventral surface of the adult hermaphrodite.

The male gonad is a single lobed, U-shaped structure which extends anteriorly, flips back on itself and then extends posteriorly to the cloaca. At the end distal to the cloaca, the germ-line nuclei are mitotic. These nuclei become meiotic, and advance through the various stages of meiosis, as they travel proximally to the seminal vesicle. The sperm are stored in the seminal vesicle until copulation, when they are released via the vas deferens and cloaca. The male tail has specialised neurons, muscles and hypodermal structures required for mating, including two spicules that are inserted into the hermaphrodite vulva to aid the transfer of sperm.

The nervous system is the most complex organ of C.elegans, accounting for 37% and 46% of somatic nuclei in hermaprodites and males, respectively (Chalfie and White, 1988). The nervous system, including all the neural connectivities, has been completely reconstructed from electron micrographs of serial sections (White et al., 1986). The system is divided into two almost independent units, the 20 nerve cells of the pharynx and the nerve cells of the rest of the body. Most of the cell bodies of these neurons are
organised into ganglia. Most of the neuronal processes from these cells form a ring around the outer surface of the pharynx (the nerve ring) or contribute to nerve bundles running the length of the body (most notably, the dorsal and ventral nerve cords). Most of the sensory neurons run anteriorly from the nerve ring to sensory structures in the head (sensilla) while most motor neurons run posteriorly from the nerve ring to the body wall muscle. A more detailed overview of *C. elegans* anatomy is given by J. White (1988).

1.2.3. *C. elegans* genetics

The haploid genome of *C. elegans* is relatively small, with approximately $10^8$ bp (100 megabases) of DNA. Wild-type hermaphrodites possess five pairs of autosomes (A) and a pair of X chromosomes (XX) while males possess five pairs of autosomes and only a single X chromosome (XO). The ratio of X chromosomes to autosomes has been shown to be the primary sex-determining signal of *C. elegans* by altering the X:A ratio (Hodgkin *et al.*, 1979; Madl and Herman, 1979).

*C. elegans* has been the focus of much classical genetic analysis and its mode of reproduction has been very useful in this respect. On one hand, genetic crosses, which are essential for classical genetic analysis, can be performed between hermaphrodites and males. On the other hand, the predominant self-fertilising hermaphroditic mode of reproduction allows easier isolation of recessive mutations, since homozygous worms appear in the F2 generation automatically, without the need for sibling crosses. Large scale genetic screens have been performed using chemical mutagenesis and radiation. Indeed, one of the earlier screens (Brenner, 1974) identified at least 77 genetic loci, distributed over the 6 linkage groups. Since then, over 1200 genetic loci have been identified using classical genetic approaches (S. Martinelli pers. comm.).

1.3. *C. elegans* as an experimental system

The features of *C. elegans* biology discussed above make it a very useful system for experimental analysis. The worm is easily maintained in the laboratory and has a short life cycle, facilitating genetic analysis. The worm is sufficiently simple to allow
thorough analysis, as evidenced by the work performed on the cell lineage and anatomy of this worm, yet is sufficiently complex to allow studies of cell interaction, animal development and behaviour. Furthermore, extensive classical genetic analysis of *C. elegans* has contributed substantially to our understanding of how individual genes participate in the development and behaviour of this nematode species.

There are several additional features which make *C. elegans* an extremely powerful experimental system. First, methods for transforming *C. elegans* have been developed which allow genetically engineered genes to be reintroduced (Stinchcomb *et al.*, 1985; Fire, 1986). Second, a physical map has been generated which covers over 95% of the *C. elegans* genome and this map has been aligned with the genetic map. The alignment of the physical and genetic maps, in conjunction with *C. elegans* transformation techniques, has greatly facilitated the isolation and cloning of mutationally defined genes using mutant rescue approaches. The physical map and its uses are discussed in more detail in Chapter 3, Sections 3.1.1-3.1.3. Third, the ongoing *C. elegans* genome sequencing project (discussed in Chapter 3, Section 3.1.2) will elucidate the structure of the entire *C. elegans* genome and will allow identification of those genes not defined by mutations. Fourth, a powerful technique for reverse genetics is available for *C. elegans* and will allow the functions of many of the genes identified by the genome sequencing project to be determined. The technique uses polymerase chain reaction (PCR) to detect insertions of the Tc1 transposon in, or near, a gene of interest. In brief, the method requires the use of mutator strains of *C. elegans* which activate Tc1 transposition and excision at high frequency in the germ-line. Genomic DNA from a number of populations of these worms are screened by PCR using primers specific to the target gene and to Tc1. Because genomic DNA is used in the PCR screen, it is necessary to use sibling selection, where siblings of the analysed animals are kept alive. Once a population of worms containing the desired Tc1 insertion has been identified, the siblings are divided into smaller sub-populations and allowed to progress through 1 generation before being analysed. By repeating the PCR screen and sibling selection steps a few times, it is possible to identify single worms containing the desired Tc1 insertion (Rushforth *et al.*, 1993). Furthermore, the ability to retrieve up to 50% viable animals from cultures of *C. elegans* stored at -70°C has permitted the construction of frozen Tc1 insertion mutant banks (Zwaal *et al.*, 1993). Such Tc1 insertion mutant banks allow the
rapid isolation of Tc1 insertion alleles in almost any gene. For example one such bank has already isolated Tc1 insertion alleles for over 70 different genes (H.C.Korswagen pers. comm.).

Antisense RNA techniques can also be used to obtain mutant phenotypes for genes identified by DNA sequencing. In *C.elegans*, inhibition of gene expression by antisense RNA has been achieved simply by injecting antisense RNA into the hermaphrodite gonad. Such experiments have demonstrated that this approach is capable of phenocopying known null mutant phenotypes of the genes being analysed. For example, antisense RNA techniques have been used to phenocopy mutations in the *par-1* maternal effect gene. These experiments demonstrated that embryos from wild-type *C.elegans* injected with *par-1* antisense RNA arrested development with a *par-1* terminal phenotype, with no morphogenesis, no intestinal cells and excess of pharynx (Guo and Kemphues, 1995). This approach has also been used to study zygotically expressed genes that act later in development. For example, antisense RNA to the CeMyoD gene (which encodes a myogenic basic helix-loop-helix protein) phenocopied a null mutation of this gene, generating embryos with a dumpy phenotype which did not develop further than the twofold stage (M.Park and M.Krause pers. comm.). Specific antisense RNA mediated gene inhibition has also been obtained by injecting *C.elegans* with constructs containing target DNA in the reverse orientation under the control of a heterologous promoter (Fire *et al.*, 1991).

The extensive knowledge of *C.elegans* biology, together with the experimental tractability of this nematode species, make it a powerful tool for isolating genes and determining gene function, using both classical and reverse genetic approaches.

1.4. *C.elegans* as a model organism

The simplicity and experimental tractability of *C.elegans*, discussed above, has facilitated the accumulation of a substantial amount of information regarding its biology. Much of this information has revealed that *C.elegans* biology shares much more in common with higher eukaryotes than might be expected from its simple body plan, supporting the contention that information obtained from this nematode species will be relevant to understanding mammalian biology. This is demonstrated by the numerous
genes isolated from \textit{C.elegans} and vertebrates which show extensive homology to one another, suggesting that \textit{C.elegans} may share many cellular pathways in common with higher eukaryotes. Two examples illustrate this well. First, several genes necessary for \textit{C.elegans} vulval differentiation encode proteins which are homologous to mammalian signal transduction proteins. The \textit{lin}-3 gene encodes an epidermal growth factor (EGF)-like molecule that is an inductive signal from the anchor cell of the gonad and which induces three ectodermal precursor cells to adopt a vulval fate (Hill and Sternberg, 1992). The \textit{let}-23 gene encodes a receptor tyrosine kinase of the EGF receptor subfamily which is thought to be the receptor for the \textit{lin}-3 product (Aroian \textit{et al.}, 1990). The \textit{let}-60 gene encodes a \textit{ras} protein that acts downstream of \textit{let}-23 to control vulval fate (Han and Sternberg, 1990; Beitel \textit{et al.}, 1990). The \textit{lin}-45 gene encodes a homologue of the mammalian \textit{raf} family of serine/threonine kinases (Han \textit{et al.}, 1993). The order of action of these genes in vulval development has been studied genetically and suggests that \textit{let}-23 positively regulates \textit{let}-60 which in turn positively regulates \textit{lin}-45. Thus, these results suggest that \textit{C.elegans} utilises a similar receptor tyrosine kinase signalling pathway for regulating vulval development as mammals do for regulating cellular responses to growth-factors.

In the second example, two genes involved in the \textit{C.elegans} programmed cell death pathway have been shown to be functional homologues of mammalian cell death genes. The \textit{C.elegans ced}-3 gene has sequence similarities to the interleukin-1\textbeta- converting enzyme (ICE). Over expression of either the murine ICE gene or \textit{ced}-3 gene in Rat-1 cells causes these cells to undergo programmed cell death (Miura \textit{et al.}, 1993). The \textit{C.elegans ced}-9 gene is required to protect cells from programmed cell death. Accordingly, loss of function mutations of this gene result in cells which should normally live, to enter the programmed cell death pathway. \textit{ced}-9 encodes a protein with sequence similarities to the mammalian \textit{bcl}-2 proto-oncogene. Over expression of \textit{bcl}-2 in \textit{C.elegans} can mimic the protective effect of \textit{ced}-9 on \textit{C.elegans} cell death and can prevent the ectopic cell deaths that occur with \textit{ced}-9 loss of function mutants (Hengartner and Horvitz, 1994). Such examples suggest that many cellular pathways may be conserved between \textit{C.elegans} and mammals and therefore support the use of \textit{C.elegans} as a general model for metazoan biology.
1.5. Cathepsin B

For several reasons, we decided to study the cathepsin B-like genes of *C. elegans*. Before discussing these reasons, I will give some background information on the current state of knowledge of cathepsin B activity and function. Cathepsin B is a proteolytic enzyme associated with the lysosomes. All proteolytic enzymes perform the same biochemical function of hydrolysing the peptide bond between two amino acid residues. Some proteases possess endopeptidase activity, cleaving peptide bonds towards the centre of a polypeptide chain, while others act as exopeptidases, cleaving bonds near the termini of a polypeptide chain. These activities are not mutually exclusive. Therefore, for the purposes of classification, all enzymes capable of acting as endopeptidases are classed as endopeptidases irrespective of whether or not they can also act as exopeptidases.

Endoproteolytic enzymes (endopeptidases) cannot be classified on the basis of substrate specificity because these enzymes do not recognise a specific substrate but rather have preferences for certain amino acid sequences in certain conformations, giving rise to a number of target substrates. Accordingly, endoproteolytic enzymes are classified according to their mechanism of catalysis. The International Union of Biochemistry currently recognises four mechanistic classes of proteolytic enzymes; the cysteine, serine, aspartic acid and metallo proteases. Accordingly, cathepsin B has been classified as a member of the cysteine protease class of endoproteolytic enzymes.

1.5.1. Papain

Papain, from *Carica papaya*, was the first recognised member of the cysteine protease class of endopeptidases. The information gained from biochemical analysis of papain was essential for the understanding of cathepsin B and therefore I will give a brief overview of the pertinent discoveries made from the studies of papain. Papain requires a free sulphydryl (-SH) group for activity and maximal activity is only obtained in the presence of thiol compounds, such as glutathione, or reducing agents, hence the terms 'thiol-dependent' or 'cysteine' are used to describe this protease, and other proteases of this class.
The amino acid sequence of the single polypeptide chain of papain was reconstructed from amino acid sequencing of peptides generated from tryptic, chymotryptic and peptic digests (Light et al., 1964). During these experiments, the use of $^{14}$C-labelled iodoacetic acid, which covalently modifies the active -SH group, demonstrated that the cysteine residue at position 25 possesses the active thiol group of papain. The tertiary structure of crystallised papain was determined by X-ray diffraction (Drent et al., 1968) and mostly agreed with the predicted primary amino acid sequence of papain. The X-ray data for papain indicates that this single 212 amino acid peptide is folded into an ellipsoid particle and possesses a binuclear structure, generated by two hydrophobic cores. The active site, and the region of interaction of the enzyme with its substrate, occurs on the surface of the papain molecule in a large groove, situated between the two hydrophobic cores of the molecule. The sulphydryl side chain of the active site cysteine residue (position 25) is found in the groove closely associated with the imidazole ring of a histidine residue (position 159). The imidazole ring is in turn hydrogen bonded to the side chain of an asparagine residue (position 175), preventing any movement of this ring. The currently accepted mechanism of catalysis by papain, and all other cysteine proteases, is shown in Figure 1.2. This mechanism is based on the X-ray structure and a substantial amount of chemical and kinetic studies performed with papain. A more detailed overview of the structure and biochemical characterisation of papain is given by Drent et al (1971) and by Glazer and Smith (1971).

Papain possesses a large catalytic site of about 25 Å which can be divided into seven ‘subsites’, each accommodating one amino acid residue of the substrate. These ‘subsites’ are positioned either side of the active site, four on one side and three on the other (Schechter and Berger, 1967). A schematic diagram of the papain active site, showing the scheme for numbering the subsites is shown in Figure 1.3. These subsites are an important basis for the substrate specificity of papain and other related cysteine protease enzymes, including cathepsin B. For example, the $S_2$ subsite of papain specifically interacts with a phenylalanine (Phe) side chain (Schechter and Berger, 1968) which results in peptides containing a Phe residue in the appropriate position having increased susceptibility to hydrolysis by papain.

As a result of the large catalytic site of papain and other related cysteine proteases, small synthetic substrates cannot accurately reflect the protein substrate
specificity of these enzymes because these synthetic substrates will only interact with a few of the subsites. Furthermore, even the use of larger peptides which interact with all the subsites will not accurately reflect the protein substrate specificity of these enzymes because recognition of such target sequences within the protein substrate will be dependent on the conformation of the protein. Thus, a protein substrate may possess several peptide sequences recognised by papain but only some, or none, will be susceptible to cleavage by this enzyme, resulting in what is called 'limited proteolysis'.

1.5.2. Isolation and biochemical characterisation of cathepsin B

The cathepsin B enzyme was first detected from bovine spleen as an activity capable of hydrolysing the synthetic substrate benzoylarginine amide in the presence of cysteine (Greenbaum and Fruton, 1957). This activity was later purified from rat liver lysosomes by chromatography and was shown to correspond to a 25 kDa protein which was named cathepsin B (Otto, 1971). Initial characterisation of cathepsin B revealed that this enzyme, and other lysosomal cysteine proteases including cathepsins L and H, show optimal activities at slightly acidic pH and are irreversibly inactivated by alkaline conditions (Barrett and Kirschke, 1981). Extensive biochemical characterisation of cathepsin B isolated from a number of vertebrates, and determination of the three dimensional structure of human liver cathepsin B, revealed that this enzyme is a cysteine protease with substantial homology to papain. Some aspects of cathepsin B revealed by these studies are discussed in the following paragraphs.

Substrate and inhibitor studies have been used extensively to characterise the activity of cathepsin B. For example, the use of synthetic substrates has demonstrated that the S2 subsite of cathepsin B has a slight preference for basic residues or phenylalanine and the S1 subsite disfavours bulky side chains (Shaw et al., 1983). Though information obtained from using small synthetic substrates cannot be used to infer the protein substrates of cathepsin B, it can be exploited in the design of substrates and inhibitors that are specific to cathepsin B. For example, the ability of cathepsin B to cleave synthetic substrates containing a pair of basic amino acid residues (Barrett and Kirschke, 1981) has proved to be a characteristic feature of this enzyme, and a variety of fluorogenic synthetic substrates which contain two arginine residues, such as Z-Arg-Arg-
NHMeC (benzyloxy carbonyl-Arg-Arg-4-methylcoumarin-7-ylamide), have been designed to take advantage of this (Barrett and Kirschke, 1981). Such synthetic substrates are very useful for discerning between different cysteine protease activities, or determining the relative activity of different cysteine proteases within a single sample.

Protein substrates have also been used in the characterisation of cathepsin B and have revealed some interesting features of cathepsin B activity. Most importantly the use of glucagon (Aronson and Barrett, 1978) and rat muscle aldolase (Bond and Barrett, 1980) as substrates has revealed that cathepsin B appears to be unique among the cysteine proteases in its ability to act both as an endopeptidase and as an exopeptidase (peptidyl dipeptidase).

Inhibitor studies have been widely used in the classification of endopeptidases because they are capable of revealing the chemical nature of the catalytic groups in a way that substrates cannot. Cathepsin B activity has been shown to be inhibited by a number of cysteine protease-specific inhibitors, indicating that it is a member of the cysteine protease class of endopeptidases. For example the cysteine protease inhibitor E-64, and derivatives of this, have been used to study inhibition of cathepsins B, H and L (Barrett et al., 1982). A large number of synthetic inhibitors have been developed which are specific for particular enzymes. These inhibitors generally comprise one or two amino acid residues linked to a reactive group. The amino acid residues of these inhibitors bind to the subsites comprising the active site and the reactive group inactivates the active thiol group of the enzyme. Thus, it is possible to obtain inhibitors which preferentially inactivate different cysteine proteases by screening those enzymes with inhibitors that possess different amino acid combinations (Kirschke and Barrett, 1987). By their nature, such inhibitors also provide information on the enzyme’s preference for certain amino acid residues and therefore provide valuable information regarding the specificity of the subsites comprising the active site of the enzyme. One such group are the peptidyl diazomethanes, some of which show very large differences in the inactivation rates for cathepsin B and cathepsin L (Shaw et al., 1983). Inhibitors specific for a particular enzyme are valuable experimental tools since they can be used to study the biological effects of inhibiting that enzyme. This approach has been used extensively for determining the biological roles of cathepsin B and related enzymes and will be discussed later.
Amino acid sequencing has elucidated the complete primary amino acid sequences of rat liver cathepsin B (Takio et al., 1983) and human liver cathepsin B (Ritonja et al., 1985) and the partial amino acid sequences of porcine liver cathepsin B (Takahashi et al., 1979; Takahashi et al., 1980) and bovine spleen cathepsin B (Pohl et al., 1982). A comparison of the primary amino acid sequences of rat and human liver cathepsin B reveals a very high level of similarity (Ritonja et al., 1985). Both enzymes are 252 amino acids long and share 83.7% identity, suggesting that they represent functional homologues. The data obtained from amino acid sequencing has demonstrated that crystallised cathepsin B from pig liver, beef spleen and rat liver is present in a mixture of single and two chain forms while human liver cathepsin B is almost exclusively in the two chain form. The two chain form is generated by limited proteolysis of the single chain form and occurs at the equivalent position for all these cathepsin B enzymes, between an asparagine and valine residue, to generate a light chain and heavy chain. These two chains are linked covalently by disulphide bonds. It is unclear whether this limited proteolysis is part of the normal processing of cathepsin B, a step in the degradation of cathepsin B, or an artefact of the purification method.

The three-dimensional structure of crystallised human liver cathepsin B has been determined (Musil et al., 1991). Comparison of the three dimensional structures of human liver cathepsin B and papain reveal that they are very similar to one another in their overall structure, including the catalytic site and active site cleft. However, the authors suggested that there were some differences which might explain the different activities of these two enzymes. Most notably, cathepsin B possesses a novel loop containing two adjacent histidine residues (His-110 and His-111) which is inserted into the active site. This loop may explain the C-terminal peptidylglycine amidating (exopeptidase) activity of cathepsin B since it is thought to favour binding of peptide substrates with two residues carboxyl-terminal to the bond which is cleaved. In such circumstances, the adjacent histidine residues of the loop are thought to anchor the carboxyl-terminal of the peptide substrate. A second difference observed is the presence of a glutamic acid residue in the S2 subsite (Glu-245) which would explain the characteristic preference of cathepsin B for basic residues at the P2 site of peptide substrates. Accordingly, mutation of the His-111 residue of recombinant rat cathepsin B to glutamine results in much reduced exopeptidase activity and increased endopeptidase activity with respect to wild-
type rat cathepsin B while mutation of Glu-245 to glutamine or alanine alters the substrate specificity of recombinant rat cathepsin B (Hasnian et al., 1992). This is a clear example of how structure relates to function. Accordingly, information obtained from the three-dimensional structure of human liver cathepsin B may be useful not only for predicting the structure of the predicted products of cloned cathepsin B-like genes for which no biochemical data is as yet available, but also for identifying residues likely to alter the activity and specificity of the predicted proteins.

1.5.3. Processing of cathepsin B

Comparison of the predicted amino acid sequence of human and rat cathepsin B (obtained by DNA sequencing of cDNA clones encoding these enzymes) to the primary amino acid sequence of mature human and rat cathepsin B (determined by amino acid sequencing) has revealed that these enzymes are translated as precursors (Chan et al., 1986; San Segundo et al., 1985). This comparison revealed that at least two cleavages must occur to release an 81 amino acid N-terminal proregion and a C-terminal hexapeptide. The predicted precursor of mouse cathepsin B, determined by sequencing of its cDNA clones (Chan et al., 1986), is very similar to those of human and rat cathepsin B, suggesting that similar processing events occur for this enzyme. Furthermore, the precursors of these three cathepsin B enzymes all possess a seventeen residue predominantly hydrophobic domain at the amino-terminus. Such signal sequences (known as prepeptides) are required to sequester the nascent protein within the lumen of the endoplasmic reticulum and are usually rapidly removed after synthesis. Thus at least two cleavage events are thought to be required to remove first the prepeptide and then the proregion from the amino terminus of these enzymes.

Recent studies using recombinant human cathepsin B expressed in yeast, suggest that procathepsin B is activated by an autocatalytic mechanism. For example, the proteolytic processing of a non-activatable form of recombinant human cathepsin B (generated by site-directed mutagenesis of the active site cysteine residue) by lysosomal enzymes in microsomal fractions is inhibited by the cathepsin B-selective CA-074 inhibitor (Mach et al., 1993), suggesting that cathepsin B activity is required for procathepsin B processing. Furthermore, activation and proteolytic maturation of human
recombinant cathepsin B has been shown to occur primarily by a concentration-independent process, indicating an intramolecular mechanism whereby procathepsin B is capable of processing itself (Mach et al., 1994a). In these experiments, a single cleavage occurred during the processing of procathepsin B to yield an intact proregion. This may be physiologically relevant since the proregion is thought to be able to act as a potent reversible inhibitor and may stabilise mature cathepsin B in vivo (Mach et al., 1994b).

These results are at odds with previous studies which suggest that cathepsin B is activated by other enzymes. For example, pepstatin a potent inhibitor of cathepsin D inhibits activation of both cathepsin B and L (Nishimura et al., 1988). Other studies, using inhibitors for each of the four classes of endoprotease, suggested that metalloprotease activity is required for processing of cathepsin B and L in macrophages (Hara et al., 1988). Such anomalies may simply reflect a more complex situation in vivo, where cathepsin B may be capable of both auto-activation and activation by other classes of enzyme. Furthermore, whilst some activation pathways may leave the proregion intact and allow it to act as an inhibitor others may degrade it completely. Thus, the type of activation pathway used may affect subsequent activity of cathepsin B.

Studies with mammalian cells have demonstrated that glycosylation is an integral part of lysosomal trafficking, with mannose-6-phosphate being the recognition signal for transport of soluble proteins to the lysosomes (see Chapter 4, Section 4.3.3.2). Accordingly, cathepsin B from human liver, rat liver and porcine spleen are all glycosylated at asparagine-111 (Ritonja et al., 1985; Takio et al., 1983; Takahashi et al., 1984). However, complete hydrolysis of human liver cathepsin B suggests that it possesses only a small carbohydrate prosthetic group compared to rat liver and porcine spleen cathepsin B (Barrett, 1977; Taniguchi et al., 1985; Takahashi et al., 1984). Furthermore, three different carbohydrate structures were identified for rat liver cathepsin B (Taniguchi et al., 1985) and two different structures for porcine cathepsin B (Takahashi et al., 1984). Whilst these two enzymes have one type of carbohydrate structure in common, none of these structures are like the high-mannose type structures normally linked to lysosomal enzymes. Together, these results suggest that vertebrates may possess different isoenzymes of cathepsin B which possess different carbohydrate moieties. Given the role of mannose-6-phosphate as a lysosomal targeting signal, these
differences may result in the targeting of different cathepsin B isoenzymes to different subcellular compartments.

1.6. Tissue distribution of cathepsin B

Cathepsin B enzymes have been isolated from human liver, rat liver, porcine spleen, bovine spleen, rabbit liver and chicken liver, suggesting that the enzyme is ubiquitous in mammals and birds (Barrett, 1977). Furthermore, cathepsin B-like activities have also been identified from a number of parasitic nematode and trematode species, as well as the free living nematode, *Caenorhabditis elegans* (discussed later), suggesting that such enzymes are also present in invertebrate metazoan species.

An immunohistochemical study of human tissues has demonstrated that cathepsin B is present in a wide variety of tissues and cell-types (Howie *et al.*, 1985). This study demonstrated that cathepsin B is present in cells of the lymphoid organs, the cardiovascular system, the respiratory system, the gastro-intestinal tract, the endocrine organs, the genitourinary system and the nervous system. Not all cells of these tissues and organs stained positive. The cell-types that frequently generated the strongest staining include macrophages, epithelial cells of various types, autonomous ganglion cells, parietal cells, kupffer cells, hepatocytes and neuronal cells. This study therefore clearly demonstrates that cathepsin B distribution is widespread if not ubiquitous throughout the human body.

1.7. The biological roles of cathepsin B

1.7.1. A lysosomal role for cathepsin B in protein turnover and degradation

Traditionally, cathepsin B has been considered to be a lysosomal enzyme because it was found to be greatly enriched in lysosomal fractions during purification (Barrett and Kirschke, 1981) and because cathepsin B shows optimal activity at acidic pH, consistent with a role in the acidic environment of the lysosomal compartment. This conclusion was supported by electron microscopy studies using a synthetic substrate specific to
cathepsin B and antisera against the substrate which demonstrated the localisation of active cathepsin B to the lysosomes of rat liver and spleen (Barrett, 1977).

Much of the early evidence of a role for cathepsin B in turnover and degradation was obtained from studying the ability of purified lysosomal enzymes to degrade purified substrates. These results suggested that cathepsin B was capable of digesting a wide variety of different protein substrates in vitro including; haemoglobin, azo-casein, azo-haemoglobin (Barrett, 1977), cartilage proteoglycans (Morrison et al., 1973) and collagen (Burleigh et al., 1974). However, in many of these studies, the physiological relevance of the conditions was not accounted for, making it impossible to determine whether these proteins were genuine substrates of cathepsin B.

In order to overcome this problem, several laboratories performed experiments in vivo using protease inhibitors. Such studies were mostly performed by labelling cellular proteins by injection of $^{14}$C-labelled leucine into rats, isolating rat livers, purifying the lysosomes and assaying the radioactivity of the products of protein degradation. Using such a system, Ahlberg et al (1985) demonstrated that lysosomes have a role in degradation of both short and long lived proteins and that leupeptin, an inhibitor of cathepsin B and other cysteine proteases, results in much reduced degradation by the lysosomes. Hopgood et al (1977) used a similar technique but isolated and cultured rat hepatocytes after $^{14}$C-leucine labelling. The authors then added protease inhibitors, including leupeptin, to the culture media to study their effect on protein degradation. The results indicated that cysteine protease activity was important for protein degradation in these cells. Shaw and Dean (Shaw and Dean, 1980) used a similar approach but with cultured mouse peritoneal macrophages, labelling protein by addition of $^{14}$C-labelled leucine into the culture media. Pepstatin was used to inhibit cathepsin D, while Z-Phe-Ala-CHN$_2$ was used to inhibit thiol proteases. Both inhibitors resulted in much reduced basal proteolysis.

Dean (Dean, 1975) studied the ability of purified cathepsin D and purified cathepsin B, as well as mixtures of lysosomal enzymes, to degrade cytosolic proteins. Again, proteins were labelled with $^{14}$C-leucine by intraperitoneal injection of rats. Cytosolic proteins were extracted from homogenised rat livers and used as a substrate. Lysosomal enzyme mixtures were obtained from purified lysosomes or lysosomal-mitochondrial fractions. Cathepsin B and cathepsin D were purified from human and rat
livers respectively. The results demonstrated that the rate of cytosolic protein degradation produced by the purified enzymes or lysosomal enzyme mixtures in vitro correlated well with rate of cytosolic protein degradation in vivo. This correlation suggested that lysosomal enzyme activity was important for cytosol protein degradation and that cathepsins B and D made significant contributions to this degradation.

As can be seen from the studies performed, much of the evidence for cathepsin B activity being involved in protein turnover and degradation is indirect because the inhibitors used in the studies are not completely specific for cathepsin B. Consequently, though there is good evidence to suggest that cathepsin B and other cysteine protease enzymes are important for intracellular protein turnover and degradation in the lysosomes, the relative input of cathepsin B to this process has not been determined.

1.7.2. A role for cathepsin B in the processing of protein precursors

Cathepsin B may process a number of protein precursors by limited proteolysis. Two pieces of circumstantial evidence support this. First, pairs of basic amino acid residues (which are highly susceptible to cathepsin B in synthetic substrates) have been observed to occur commonly at the point of cleavage of activation peptides from the precursors of polypeptide hormones and proteases (Steiner et al., 1974). For example, the first steps of proinsulin processing seem to be cleavage at two sites which have this type of primary structure (Chance, 1972). Second, cathepsin B activity has been identified in organelles where prohormone processing is thought to occur. For example, it has been found in the purified secretion granules of a rat insulinoma and in normal rat islet granule fractions (Docherty et al., 1983).

More recently, evidence of a role for cathepsin B in precursor processing comes from its own biosynthesis pathway, where procathepsin B has been shown to be capable of autoactivation by cleavage of the proregion (Mach et al., 1993; Mach et al., 1994a). There is also evidence to suggest that cathepsin B in conjunction with other lysosomal proteases, is responsible for the processing of thyroglobulin into active thyroxin molecules. First, thyroglobulin processing in follicular cells is thought to occur in phagolysosomes generated by the fusion of lysosomes with vesicles which transport thyroglobulin taken up from the intra-follicular space. Immunogold electron microscopy
has been used to demonstrate that cathepsin B, H and L are all co-localised with Thyroxin (T4) in the colloid droplets proximal to the apical membrane within these cells (Kominami and Uchiyama, 1993). Second, selective inhibition of lysosomal proteases from purified lysosomes results in a significant reduction in thyroglobulin degradation (Dunn et al., 1991b). Third, in vitro proteolysis of in vivo $^{125}$I-labelled rabbit thyroglobulin by cathepsin B, L and D has shown that these enzymes are able to cleave thyroglobulin at specific sites, suggesting a direct role in processing of this hormone precursor (Dunn et al., 1991a).

Cathepsin B is also thought to be the major renin processing enzyme. Crude homogenates of human kidney have been shown to contain a thiol protease which cleaves prorenin at the same site clipped in native renin (Shinagawa et al., 1990). Removal of cathepsin B from such homogenates, using anti-cathepsin B antibodies or the cathepsin B specific inhibitor CA-074 abolishes this renin-processing activity (Kominami and Uchiyama, 1993). Furthermore, the authors demonstrated the co-localisation of prorenin and renin with cathepsin B in immature secretory granules using electron microscopy.

1.7.3. A role for cathepsin B in antigen processing

There is a substantial amount of evidence to suggest that cathepsin B performs important roles in the immune system. It is thought to be involved, with other lysosomal enzymes, in the digestion of endocytosed antigens for display by MHC class II molecules on the surface of antigen presenting cells (APCs). Electron microscopy studies have demonstrated the co-localisation of active cathepsin B and cathepsin D with MHC class II-I chain complexes in the early endocytic compartment, where contact with endocytosed antigens is most likely (Guagliardi et al., 1990). Furthermore, processing of certain antigens have been shown to require the activity of cathepsin B in vivo. The immune responses of mice to vaccines of hepatitis B surface antigen (HBsAg) and rabies are inhibited by specific inhibitors of cathepsin B, specific synthetic substrates of cathepsin B or anti-cathepsin B F(ab)' antibody fragments (Matsunaga et al., 1993). In addition, using splenocytes primed with a synthetic epitope predicted to be a product of HBsAg cleavage by cathepsin B, these authors demonstrated that rechallenge with a
mixture of native HBsAg and cathepsin B inhibitors produced only a weak proliferative response while rechallenge with a mixture of the synthetic epitope and cathepsin B inhibitors produced a strong proliferative response similar to native HBsAg alone. The authors concluded that this epitope was a powerful antigenic peptide and was naturally produced by cathepsin B in vivo.

The I chain regulates presentation of antigen by MHC class II molecules in several ways. It associates with MHC class II α,β chains during synthesis in the endoplasmic reticulum. Subsequently, it regulates intracellular transport of MHC class II α,β chains through cytoplasmic signals. Finally, it inhibits binding of endogenously derived peptides to MHC class II α,β chains in living cells. Cathepsin B has been shown to digest the I chain in a staged pattern suggesting a mechanism for regulating the binding of antigenic peptides to MHC class II. (Xu et al., 1994). In support of this, evidence has been obtained that cysteine proteases cleave the I chain in vivo. Morton et al (1995) studied the trafficking, proteolysis and dissociation of I chain associated with nascent MHC class II molecules in B-lymphoblastoid cells using metabolic labelling. These results demonstrated that the MHC class II-I chain complexes were catabolised rapidly after entry into lysosome-like compartments but that this catabolism was blocked by cysteine protease inhibitors.

Interestingly, gamma interferon (γ-IFN), which is known to stimulate the presentation of peptides and expression of MHC class II molecules by APCs such as macrophages and lymphocytes, has been shown to cause a selective induction of cathepsins B and L in macrophages (Lah et al., 1995). The authors suggest that γ-IFN may be responsible for upregulating all genes necessary for the display of peptides by MHC class II molecules, and that cathepsins B and L are part of this pathway.

1.7.4. A role for cathepsin B in bone remodelling

Bone resorption occurs beneath the ruffled border of the mammalian osteoclast in a sealed extracellular microenvironment that resembles a phagolysosome in its acidic pH. The calcified bone is resorbed by two processes. First, acidification, which requires carbonic anhydrase activity and proton pumps, is responsible for solubilising the mineral portion of bone. Second, the organic matrix of the bone (mostly collagen) is degraded
by a range of enzymes including lysosomal cysteine proteases, lysosomal hydrolases and collagenases. A role for cysteine proteases in bone resorption was first suggested after lysosomal enzymes were found to be able to degrade protein components of the extracellular matrix (see Section 1.7.1). Direct evidence for the involvement of cysteine proteases in bone resorption was subsequently obtained from experiments using specific enzyme inhibitors. Cysteine protease inhibitors were demonstrated to inhibit bone resorption by *in vitro* cultured mouse calvaria (Delaisse et al., 1980). Subsequently, cysteine protease inhibitors were shown to inhibit bone resorption both *in vitro* and *in vivo* (Delaisse et al., 1984), using cultured mouse calvaria and male wistar rats.

Once cysteine proteases had been implicated in bone resorption, research was directed towards determining which of the cysteine proteases might be important and how they might exert their effect. Both cathepsins B and L activities have been detected in protein extracts of bone tissue and cathepsin B activity has been detected in the media of cultured mouse calvaria (Delaisse et al., 1991) suggesting these enzymes may be important effectors of bone resorption. However, cathepsin L has been demonstrated to degrade collagen and gelatin substrates at a much higher rate than cathepsin B (Delaisse et al., 1991), suggesting that cathepsin B may not be a direct effector of bone resorption but rather may have a role in activating other enzymes important for this process. Recently, Hill et al (1994) obtained evidence for a primarily intracellular role for cathepsin B. The authors used membrane permeable and impermeable forms of two inhibitors of cysteine proteases: Ep475 which inhibits cathepsins B, H, L and S and is unable to traverse membranes; Ep453, the ethyl ester of Ep475, which has the same specificity as Ep475 and is membrane permeable; CA-074, which specifically inhibits cathepsin B and is unable to traverse membranes; CA-074Me, the methyl ester of CA-074, which has the same specificity as CA-074 and is membrane permeable. The authors demonstrated that while Ep475, Ep453 and CA-074Me all inhibited bone resorption *in vivo* and *in vitro*, the membrane impermeable cathepsin B-specific inhibitor CA-074 had little effect on bone resorption. These results strongly suggest that while cathepsins L and S act extracellularly, and possibly intracellularly, cathepsin B acts mostly intracellularly. Accordingly, the authors suggested that cathepsin B may have a primary role in activating enzymes involved in collagen degradation such as other cysteine proteases or matrix metalloproteases.
1.7.5. A pathological role for cathepsin B in tumour cell invasion

Since biochemical studies (see Section 1.7.1) first demonstrated the ability of cathepsin B to degrade components of the extracellular matrix, this enzyme has attracted a substantial amount of work in the cancer research field because of its potential role as an effector of tumour cell invasion. These studies have resulted in the accumulation of a substantial amount of indirect evidence that suggests cathepsin B may be involved in tumour cell invasion. Elevated levels of cathepsin B activity have been correlated with increased metastatic potential or invasiveness in a variety of tumours from humans including; colorectal carcinomas (Sheahan et al., 1989), neoplastic cervical cells (Pietras and Roberts, 1981) and adenocarcinomas (Recklies et al., 1980). Such correlations have also been observed with tumours from rodents. For example; elevated cathepsin B activity has been observed in murine melanomas (Sloane et al., 1982), while elevated cathepsin B mRNA transcript levels have been observed in murine melanomas (Qian et al., 1989) and hepatomas (Moin et al., 1989).

Cathepsin B from tumour cells may not only be expressed at elevated levels but also expressed as forms having larger size and/or altered stability when compared to normal tissues. For example, a study of osteoclastoma tissue revealed six cysteine protease activities (Page et al., 1992). The six activities were all more similar to cathepsin B than other lysosomal cysteine proteases. All showed kinetics of inhibition by synthetic diazomethane inhibitors as expected for cathepsin B and all showed reduced activity to native collagen. Furthermore, antibodies against cathepsin B reacted against all six activities. Northern blotting using a cathepsin B cDNA probe also revealed three species of transcript for this enzyme. These results suggested that multiple forms of cathepsin B were being expressed by osteoclastomas, some of which were distinct from those observed in normal tissues.

There is a substantial amount of evidence to suggest that cathepsin B-like enzymes are secreted. Cathepsin B-like enzymes which react with anti-cathepsin B antibody and show the characteristic synthetic substrate preferences of cathepsin B have been detected in the ascitic fluid of patients with neoplasia (Mort et al., 1983) and in the pleural effusions of breast cancer patients (Petrova-Skalkova et al., 1987). Cathepsin B
has also been detected in media conditioned by tumour cells, or explants, from a variety of different sources. Using anti-cathepsin B antibodies, cathepsin B-like enzymes have been detected in the media of cultured malignant human colorectal carcinoma cell lines (Maciewicz et al., 1989) and in the media of cultured human breast tumours (Recklies et al., 1982b). Using biochemical analysis, enzymes with synthetic substrate preferences and enzyme kinetics characteristic of cathepsin B have been detected in the media of cultured malignant human breast tumours (Mort et al., 1980) and in the media of cultured rabbit V2 carcinoma cells (Baici and Knopfel, 1986).

In addition to being secreted, cathepsin B has also been detected in unusual cellular compartments of tumour cells. The use of synthetic substrates and synthetic inhibitors in cell fractionation experiments has demonstrated a cathepsin B-like enzyme activity on the cell surface associated with the plasma membrane of several murine melanoma and fibrosarcoma cell lines (Keren and LeGrue, 1988). Similar experiments have also identified cathepsin B-like activity strongly bound to the plasma membrane of human neoplastic cervical cancer cells (Pietras and Roberts, 1981).

Despite the large amount of indirect evidence available, there has been much debate as to the significance of the presence of membrane-associated or secreted forms of cathepsin B in tumour cells. This debate has largely arisen from two observations. First, a number of non-tumour cell lines have been shown to secrete a proportion of their cathepsin B, indicating that small amounts of cathepsin B may be present in the extracellular environment in non-disease states. For example, activatable precursor forms of cathepsin B are secreted not only by human lung tumours of different histological cell types but also by normal fibroblasts and alveolar macrophages, indicating that this process cannot be considered tumour cell specific (Werle et al., 1994).

Second, most tumour cells secrete the majority of cathepsin B in an inactive precursor form. Indeed, cathepsin B activity in media from cultured tumour cells rarely shows much cathepsin B activity unless it is activated by an enzyme such as pepsin (Mort et al., 1983; Baici and Knopfel, 1986; Petrova-Skalkova et al., 1987; Keppler et al., 1988; Werle et al., 1994), suggesting that secretion of cathepsin B may not be pathologically significant.

In answer to this, many researchers have suggested that the enhanced stability at neutral and/or alkaline pH observed for a number of secreted cathepsin B-like enzymes
may allow such enzymes to be active in conditions mimicking the extracellular milieu (Mort et al., 1980; Recklies et al., 1982a; Baici and Knopfel, 1986; Keppler et al., 1988), in contrast to lysosomal cathepsin B which is known to be rapidly inactivated under alkaline conditions. More recently, Mach et al. (1994b) have expressed recombinant human cathepsin B in yeast and isolated an active, high molecular weight form of recombinant human cathepsin B that is stable at neutral or slightly alkaline pH. This high molecular mass form was found to be a non-covalent complex between the mature recombinant cathepsin B and its propeptide. This active form was capable of hydrolysing synthetic substrates, but it was unable to hydrolyse protein substrates. However, incubation at acid pH resulted in liberation of mature active cathepsin B enzyme. This is in agreement with other studies which have demonstrated that the synthetic rat cathepsin B proregion acts as a powerful reversible inhibitor of the enzyme (Fox et al., 1992), but that the inhibitory effect of this proregion is pH-dependent since acidification of the conditions causes degradation of the proregion by the enzyme itself. The authors suggested that this high molecular mass, non-covalent complex represented an important mechanism by which extracellular cathepsin B could remain dormant until local acidification allowed release of the mature active enzyme. Thus, it is possible that the cathepsin B-like enzymes secreted by some tumour cells may be activatable in the extracellular environment and may therefore be important effectors in tumour cell invasion.

1.8. Cathepsin B and parasitism

There is an increasing amount of evidence to suggest that parasitic nematode and trematode species may require cysteine proteases, including cathepsin B-like enzymes, for parasite-specific processes including parasite metabolism and host invasion.

There is a substantial amount of evidence to suggest that a cathepsin B-like enzyme is essential for Schistosoma mansoni metabolism. Schistosomes feed on red blood cells. These cells are lysed within the digestive tract of the parasite and haemoglobin is released. The haemoglobin is degraded in the gut producing a dark pigment end-product termed haematin. Haemoglobin digestion is therefore thought to be the principle mechanism by which schistosomes obtain essential amino acids. Two
highly immunogenic proteins, named Sm31 and Sm32, have been identified in extracts from *S. mansoni* and immunofluorescence studies have localised these proteins to the cells of the digestive tract (Ruppel *et al.*., 1987). Two clones encoding Sm31 and Sm32 have been isolated from an *S. mansoni* cDNA expression library, using antisera from infected mice and humans (Klinkert *et al.*, 1987). Sequence analysis of these clones has revealed that Sm31 shares significant homology to cathepsin B. Furthermore, the expression of Sm31 in insect cells has generated a proteolytically active product capable of cleaving haemoglobin (Gotz and Klinkert, 1993). These observations suggest that the cathepsin B-like Sm31 enzyme is the 'haemoglobinase' of *S. mansoni*.

Cathepsin B-like enzymes may also perform important metabolic functions in two other trematode species, *Fasciola Hepatica* (McGinty *et al.*, 1993) and *Haplometra cylindracea* (Hawthorne *et al.*, 1993). In both cases, cathepsin B-like activities were detected in the culture media (using active site-directed affinity labels and synthetic substrates), suggesting that these enzymes may be required for digestion by these trematode species. For *F. hepatica*, a role in digestion for such enzymes is supported by the detection of cathepsin B-like enzyme activity in the secretory granules of *F. hepatica* intestinal cells (Yamasaki *et al.*, 1992). For *H. cylindracea*, which feeds exclusively on blood, the cathepsin B-like activity detected in the culture media may also be important for digestion since it was able to efficiently degrade haemoglobin (Hawthorne *et al.*, 1993).

Many parasites penetrate and migrate through host tissues during certain stages of their life cycles. It is generally believed that this must involve proteolytic degradation of the extracellular matrix. Early histochemical studies on tissue invasion by larval stages of the parasitic nematode species *Strongyloides ratti*, *Strongyloides simiae*, *Nippostrongylus muris* and *Trichinella spiralis* and the parasitic trematode species *Schistosoma mansoni* and *Schistosomatium douthitti* (Lewert and Lee, 1954) indicated disruption of host tissues around the paths of migration. The authors noted that the substantial alteration of the basement membrane and ground substances of the hosts' connective tissue was consistent with proteolytic degradation by enzymes secreted by the infective larvae. This was supported by the detection of collagenase activity from the larvae of many of the parasitic nematode and trematode species tested.
Since the early observations of Lewert and Lee, a substantial amount of indirect evidence has accumulated to suggest that many parasitic nematode and trematode species secrete proteolytic enzymes to facilitate migration through host tissues, and that some of these may be cathepsin B-like enzymes. This evidence has been obtained from the analysis of media conditioned by culture of parasitic nematode and trematode species in vitro. The principle problem of this approach is that it is not possible to determine whether proteolytic enzymes detected in the media have been released by active processes, such as secretion or excretion, or are a result of release of proteins after death of the parasites, which may occur as a consequence of in vitro culture. However, this problem has largely been resolved for a number of parasitic nematode and trematode species by reducing the in vitro culture time, by improving culture conditions to increase viability and by checking for worm viability prior to removing samples of the culture media. The advances made with in vitro culture has resulted in acceptance of the term ‘Excretory / Secretory’ (E/S) products to describe proteins detected in media conditioned by parasitic nematode and trematode species. However, some authors refer to these as ‘In Vitro Released’ (IVR) products in recognition of their uncertain origins. It is important to note that, though E/S products detected in the media are probably derived from active processes, the nature of the active processes that give rise to these products cannot be determined in such studies. Thus, E/S products detected in the culture media may be released as a result of secretion from glands, as a result of excretion from the alimentary canal, or as a result of regurgitation of enzymes from the mouth parts. Accordingly, these types of study alone cannot distinguish whether the E/S products detected in culture media may have primary roles outside the worm, suggestive of a role in host invasion or immune evasion, or primary roles within the worm, suggestive of a role in digestion.

At least one cathepsin B-like enzyme appears to be secreted from the nematode hookworm Ancyclostoma caninum (Harrop et al., 1995). The authors isolated two cDNA clones which, though distinct from one another, both showed significant homology to human cathepsin B. The authors expressed one of these cDNA clones in E.coli and raised rabbit antisera to the resulting AcCP-1 recombinant protein. The antisera reacted with a single 40kDa antigen from E/S products of A.caninum. Immunohistochemical studies of serial sections of A.caninum revealed that the antisera
reacted with esophageal, amphidial and excretory glands. These data provide strong evidence that a cathepsin B-like enzyme is secreted by this nematode species.

Analysis of the E/S products of the parasitic trematode *F. hepatica* using active-site-directed affinity labels has identified a major cathepsin B-like cysteine protease from both juvenile and adult worms (McGinty *et al.*, 1993). The identification of this activity in the juvenile stages, which exhibit tissue invasive behavior, suggests that this enzyme may be involved in host invasion or tissue migration. A role in tissue migration is supported by the observation that the cathepsin B-like activity detected showed unusual stability at alkaline pH and might therefore function in the extracellular environment.

The use of specific inhibitors has revealed the presence of cysteine protease activities in both the exsheathing fluid E/F and E/S products from the L₃ infective stage of the nematode *Necator americanus* (Kumar and Pritchard, 1992). Interestingly, a single cysteine protease activity was defined in the E/F fluid, but not from somatic extracts of L₃ larvae, suggesting that a single cysteine protease enzyme appears only during exsheathment of the larvae and may therefore perform a role in ecdysis. In contrast, the E/S products contained several activities that were attributable to cysteine protease enzymes, and therefore might have a role in tissue degradation during host invasion. Since only general cysteine protease inhibitors were used, it is not possible to determine whether cathepsin B-like enzymes contributed to the cysteine protease activity observed.

A cysteine protease enzyme of 28 kDa has also been partially purified from both E/S fluids and whole worm homogenates of the trematode *Spirometra mansoni* (Yong Song and Chappell, 1993). The partially purified cysteine proteases isolated from worm homogenates and from E/S fluids have identical pH profiles, both cleave collagen to produce the same products and both show negligible ability to cleave haemoglobin. Together, these data suggest that this enzyme is active both within the worm and outwith the worm. The authors concluded that the enzyme may be involved in host migration because of its ability to cleave collagen. However, a role for this enzyme in digestion is also possible since active enzyme was also present within the worm.

The E/S products generated by culture of L₂ and L₃/₄ stages of the parasitic nematode *Ascaris suum* have also been studied (Knox and Kennedy, 1988). These experiments detected proteolytic activities of all four mechanistic classes. Furthermore,
inhibitors specific to each of the four classes had different effects on the overall proteolysis of the selected protein substrates dependent on the pH and on which of the two life cycle stages were studied. These data suggest that both larval stages require the activity of more than one proteolytic enzyme and that these enzymes are required in a stage specific manner. A requirement for secreted enzymes from multiple mechanistic classes is probably common to all parasitic nematode and trematode species since a single enzyme class is unlikely to be able to fulfil all the biological functions performed by secreted enzymes.

Cysteine protease activities have also been identified in homogenates of a number of nematode species. For example, analysis of the E/S fluids and whole worm homogenates of both L₃ and L₄ larvae of *Dirofilaria immitis* using inhibitors to metallo and cysteine proteases has revealed that cysteine protease activity is only present in the homogenates (Richer *et al.*, 1992). Analysis of homogenates of *Strongyloides ransomi* has identified two cysteine proteases with molecular weights of 32kDa and 28kDa which show pH optima and substrate preferences similar to vertebrate cysteine proteases (Dresden *et al.*, 1985).

These studies highlight how limited our knowledge is of parasitic nematode and trematode cysteine protease biology when compared to vertebrates. The relative lack of knowledge reflects a problem faced by all parasitologists, namely the difficulties associated with growth and maintenance of parasites in the laboratory. Such difficulties limit the amount of material available for analysis and therefore hamper detailed biochemical analysis. Accordingly much of the biochemical analysis of parasite cysteine proteases in protein extracts is limited to detection and basic characterisation of activity. Despite these problems, the study of cysteine proteases in parasitic nematode and trematode species has demonstrated that some may secrete and/or excrete enzymes with cathepsin B-like activities. In turn, this suggests that the general cysteine protease activities identified in the E/S products of other parasitic nematode and trematode species may also be due in part to the presence of cathepsin B-like enzymes. These studies have also provided sufficient information to suggest that cysteine proteases, including cathepsin B-like enzymes, are used by parasitic nematode and trematode species for very different biological roles than the lysosomal cysteine proteases of vertebrates.
1.9. Cathepsin B-like multigene families

The problems associated with the study of parasites (discussed above) have resulted in the use of molecular biology as a tool for studying parasite cysteine protease biology. Accordingly, cysteine protease genes have been cloned from several unicellular and multicellular parasite species. However, to date, cathepsin B-like multigene families with the potential to encode cathepsin B-like enzymes have only been reported in parasitic nematode and trematode species. The parasitic nematode *Haemonchus contortus* possesses at least five cathepsin B-like genes (Cox *et al.*, 1990; Pratt *et al.*, 1990; Pratt *et al.*, 1992a) while the cathepsin B-like multigene family of the parasitic nematode *Ostertagia ostertagi* is thought to comprise at least two members (Pratt *et al.*, 1992b). In addition, the isolation of two cDNA clones from *Ancylostoma caninum* that encode distinct enzymes with homology to human cathepsin B suggests this parasitic nematode species also possesses a cathepsin B-like multigene family (Harrop *et al.*, 1995). There is also some evidence for a cathepsin B-like multigene family in the parasitic trematode *Fasciola hepatica* (Heussler and Dobbelraere, 1994). The authors used polymerase chain reaction to amplify seven different cDNA clones. Five of these clones showed homology to human cathepsin L and two showed homology to *Schistosoma mansoni* Sm31 encoding a cathepsin B-like gene. However, the two cDNA clones with homology to cathepsin B were truncated in the region of the active cysteine residue and therefore it is not possible to confirm that these clones encode cathepsin B-like enzymes. The identification of cathepsin B-like multigene families in parasitic nematode species, and possibly also parasitic trematode species, suggest that this class of enzyme is important for the biology of these species. This is in agreement with the biochemical studies discussed in Section 1.8 which suggest that cathepsin B-like enzymes may be required for parasite-specific processes in some parasitic nematode and trematode species.
1.10. **Cathepsin B-like enzymes in *C.elegans***

Two similar but distinct cysteine protease activities have been detected in whole worm extracts of the free living nematode species *Caenorhabditis elegans* and have been named Ce1 and Ce2 (Sarkis et al., 1988b). These two activities cannot be distinguished on the basis of specificity for model substrates, cofactor requirements or inhibitor sensitivities. Both cathepsins Ce1 and Ce2 are thiol-dependent, and are both inhibited by the cysteine protease inhibitors leupeptin and E-64. They both have the same pH optima of 5.0 for the same fluorogenic synthetic substrate Z-Phe-Arg-MCA, indeed both enzymes showed maximal activity with this substrate, which is known to be susceptible to both vertebrate cathepsins B and L. However, both these enzymes possess features intermediate between cathepsins B and L. Neither of these enzymes hydrolyse Z-Arg-Arg-MCA, a characteristic substrate for cathepsin B, suggesting that these two enzymes may be more cathepsin L-like. Conversely, both enzymes display peptidyl-dipeptidase activity, a feature thought to be unique to cathepsin B among the papain-like cysteine protease enzymes. Thus Ce1 and Ce2 both appear to be cathepsin B-like and cathepsin L-like in their activities. These two enzymes have been shown to normally reside in lysosomes by Sarkis et al (1988a) which is unsurprising in view of their acidic pH optima and similarity to the lysosomal cathepsins B and L. A third cysteine protease activity, named Ce3, was also detected which was very different to Ce1 and Ce2, and indeed to vertebrate cathepsins B and L. The enzyme exhibited optimal activity at pH 5.5 and was thiol dependent, yet was also leupeptin insensitive (unlike Ce1 and Ce2). This enzyme showed only very weak activity towards Z-Phe-Arg-MCA and would not hydrolyse any of the other synthetic or peptide substrates used in the study.

A gene encoding a cathepsin B-like enzyme has been cloned from *C.elegans* and named *cpr-1* (Ray and McKerrow, 1992). The predicted protein encoded by this gene appears to share the same preproenzyme structure as human cathepsin B, and shares almost 70% similarity in the mature enzyme region. The similarity of *cpr-1* to vertebrate cathepsin B suggests that these two enzymes will have a very similar structure.

However, *in-situ* hybridisation experiments suggest that *cpr-1* is expressed only in the intestinal cells lining the gut of *C.elegans*. Such tissue restricted expression is quite unlike the expression patterns observed for vertebrate cathepsin B which is thought
to be expressed in most tissues of the body (Howie et al., 1985). Furthermore, partial sequencing of randomly selected cDNA clones from a partially normalised C.elegans cDNA library identified several clones encoding partial protein sequences with homology to human cathepsin B (Waterston et al., 1992). This suggests that C.elegans may, like the parasitic nematodes H.contortus and O.ostertagi, possess a cathepsin B-like multigene family.

1.11. Project Aims and Objectives

As a result of the extensive characterisation of vertebrate cathepsin B, the structure, function, distribution and biological roles of this enzyme are well understood. In contrast, there is only a limited amount of information available for the cathepsin B-like enzymes of parasitic nematode and trematode species. Despite this imbalance, there is sufficient evidence to suggest that cathepsin B-like enzymes perform different biological functions in parasitic nematode and trematode species to the functions performed by cathepsin B in vertebrates. These differences may make the cathepsin B-like enzymes of parasitic nematode and trematode species useful targets for serodiagnostic and/or chemotherapeutic agents. However, the effective design and application of such agents requires a sufficient knowledge of basic parasitic nematode and trematode biology and this knowledge is proving hard to obtain.

The evidence obtained to date suggests that C.elegans may encode a cathepsin B-like multigene family and that one member of this family exhibits gut-specific expression. These data suggest that the cathepsin B-like enzymes of C.elegans, like those of parasitic nematode and trematode species, may perform very different biological functions to the cathepsin B enzyme of vertebrates. Accordingly, I have initiated a study of the cathepsin B-like genes in this free living nematode species. C.elegans is well characterised and is a powerful tool for the isolation and analysis of new genes. Thus, the aim of this project was to isolate and characterise the cathepsin B-like genes of C.elegans and to gain an insight into the potential biological roles of the enzymes encoded by these genes. The information gained from these studies would not only generate more understanding of basic nematode biology but might also provide valuable insights into the function of cathepsin B-like enzymes in parasitic nematode species.
Figure 1.1

A  A diagram of a cross section through the adult hermaphrodite, taken through the posterior end and viewed towards the posterior (Edwards and Wood, 1983).

Key:  m, muscle; h, hypodermis; i, intestine; n, nerve cord; g, gonad.

B  Schematic diagrams showing the lateral view of a mature adult hermaphrodite and a mature adult male (Sulston and Horvitz, 1977).
**Adult Hermaphrodite**

- pharynx
- Intestine
- ovary
- terminal bulb of pharynx
- uterus
- vulva
- eggs
- spermatheca
- oocytes

**Adult Male**

- Intestine
- pharynx
- testis
- vas deferens
- terminal bulb of pharynx
- copulatory apparatus
Figure 1.2

Schematic representation of the steps involved in catalysis by members of the cysteine protease class of enzyme. This diagram shows only the reaction through to the acyl enzyme intermediate. Breakdown of this intermediate involves an enzyme-catalysed attack of water (Dunn, 1989).
A schematic diagram of papain showing the Schechter and Berger nomenclature for binding of a peptide substrate to papain. The seven subsites of papain are labelled $S_1$-$S_4$ and $S_1'$-$S_4'$ and the amino acid residues of the peptide substrate corresponding to these sites are labelled $P_1$-$P_4$ and $P_1'$-$P_4'$. The peptide bond hydrolysed is also indicated.

The diagram was adapted from Schechter and Berger (1967).
Chapter 2
Chapter 2

Materials and Methods

Unless otherwise stated, all chemicals / reagents were purchased from BDH (Merck House, Poole, Dorset, UK), GibcoBRL (Life Technologies Ltd., Paisley, Renfrewshire, UK) or Sigma (Sigma-Aldrich Company Ltd., Poole, Dorset, UK). Restriction enzymes were obtained from GibcoBRL (Life Technologies Ltd.).

2.1. Chemical abbreviations

BSA: bovine serum albumen
DEPC: diethyl pyrocarbonate
diH₂O: deionized water
IPTG: isopropyl-β-D-thiogalactoside
MOPS: 4-morpholinepropanesulphonic acid
Na₂EDTA: ethylenediaminetetraacetic acid disodium salt
PEG-8000: polyethylene glycol, molecular weight approximately 8,000
sdiiH₂O: sterile deionized water
SDS: sodium dodecyl sulphate
Taurine: 2-aminoethanesulphonic acid
TEMED: N₁N₁N₁N₁-tetramethylethylene diamine
Tris-HCl: 2-amino-2-(hydroxymethyl)-1,3-propanediol-hydrochloride
X-gal: 5-bromo-4-chloro-3-indolyl-β-D-galactopyranoside (Boehringer Mannheim)

2.2. Commonly used stocks, solutions and media

Ampicillin (1,000x): 100mg/ml solution in sdiiH₂O. Aliquoted and stored at -20°C.

BBL Agar: 0.6% Agar, 1% Trypticase BBL (Becton Dickinson, Oxford, UK), 0.5% NaCl, 10mM MgSO₄ prepared with diH₂O. Autoclaved and stored at room temperature.
**BBL Agarose:** 1% Trypticase BBL (Becton Dickinson), 0.5% NaCl, pH 7.2 with NaOH, 0.65% Agarose, 0.25% MgSO₄ prepared with diH₂O. Autoclaved and stored at room temperature.

**Denaturation Buffer:** 1.5M NaCl, 0.5M NaOH in diH₂O. Stored at room temperature.

**Denhardt's Solution (50x):** 1% BSA, 1% Ficoll, 1% polyvinyl pyrrolidone in sdiH₂O. Aliquoted and stored at -20°C.

**DEPC Treated ddH₂O:** DEPC added to diH₂O to a final concentration of 0.1%, mixed vigorously for 10 min and autoclaved. Stored at room temperature.

**Ethidium Bromide:** 10mg/ml stock in diH₂O, final concentration 0.5μg/ml in agarose gels. Stored, protected from light, at room temperature.

**Herring Testes DNA:** 10mg/ml stock prepared in sdiH₂O and sheared using a 19 gauge hypodermic needle and luer-lock syringe (Becton Dickinson). Stored at -20°C.

**IPTG:** 0.5M stock in sdiH₂O. Aliquoted and stored at -20°C.

**Kanamycin (1,000x):** 50mg/ml in sdiH₂O. Aliquoted and stored at -20°C.

**L-Broth:** 1% Bacto tryptone (Difco, Michigan, USA), 0.5% yeast extract (Difco), 0.5% NaCl, in diH₂O. Autoclaved and stored at room temperature.

**L-Broth Agar:** as for L-Broth + 1.5g/100ml bacto-agar (Difco). Autoclaved then allowed to cool (antibiotics were added at this stage, if required) poured with sterile technique on to plastic plates, allowed to set, dried by inversion in 37°C oven and stored, for short periods, at 4°C.
M9 Buffer (1x): 0.3% KH$_2$PO$_4$, 0.6% Na$_2$HPO$_4$, 0.5% NaCl, 1mM MgSO$_4$, in diH$_2$O. Autoclaved and stored at room temperature.

MOPS Buffer (10x): 0.2M MOPS Buffer, 0.05M sodium acetate, 0.01M Na$_2$EDTA, in sdiH$_2$O. Stored at 4°C.

Na$_2$EDTA: 0.5M stock in diH$_2$O pH8.0 with NaOH. Autoclaved and stored at room temperature.

Neutralisation Buffer: 1.5M NaCl, 0.5M Tris-HCl pH8.0 in diH$_2$O. Stored at room temperature.

NGM Agar: 0.3% NaCl, 1.7% agar (Difco), 0.25% peptone (Difco), 0.0003% cholesterol (1ml/L of 5mg/ml stock in ethanol), in diH$_2$O. Autoclaved and then 1ml/L of 1M CaCl$_2$, 1ml/L of 1M MgSO$_4$ and 25ml/L of 1M potassium phosphate buffer pH6.0 added. Stored at room temperature.

Phenol: Purchased liquefied and pre-equilibrated. Stored at 4°C.

Phenol / Chloroform: Mix of equal volumes of phenol and chloroform. Stored at 4°C.

Potassium Phosphate Buffer (1M): Prepared from 1M KH$_2$PO$_4$ and 1M K$_2$HPO$_4$, mixed in appropriate ratio for desired pH. Autoclaved and stored at room temperature.

SDS: 10% stock solution in sdiH$_2$O. Stored at room temperature

Sodium Phosphate Buffer (1M): Prepared from 1M NaH$_2$PO$_4$ and 1M Na$_2$HPO$_4$, mixed in appropriate ratio for desired pH. Autoclaved and stored at room temperature.

SM Buffer: 50mM Tris-HCL (pH7.5), 0.1M NaCl, 8mM MgSO$_4$, 0.01% gelatin, in diH$_2$O. Autoclaved and stored at room temperature.
SSC (20x): 3M NaCl, 0.3M tri-sodium citrate. pH7.0 with NaOH, in diH₂O. Stored at room temperature.

SSPE (20x): 3.6M NaCl, 0.2M NaH₂PO₄, 0.02M Na₂EDTA pH8.0 with NaOH in diH₂O. Stored at room temperature.

TE: 1mM EDTA, 10mM Tris-HCl, (pH as required) in diH₂O. Autoclaved and stored at room temperature. TE at pH 8.0 was used for all DNA manipulations, unless otherwise indicated.

Tetracycline (1,000x): 12mg/ml in 50% sdiH₂O, 50% ethanol. Aliquoted and stored at 4°C.

TBE (10x): 0.9M Tris-HCl, 0.9M Boric acid, 25mM Na₂EDTA pH8.0, in diH₂O. Stored at room temperature.

TTE (20x): 1.8M Tris-HCl, 0.6M Taurine, 0.01M Na₂EDTA, in diH₂O. Stored at room temperature.

X-Gal (100x): 2% in dimethyl formamide (DMF). Stored in aliquots, protected from the light, at -20°C.

YPD Broth: 2.0% glucose, 2.0% bacto peptone (Difco) and 1% yeast extract (Difco) in diH₂O. Autoclaved and stored at room temperature.

YCD Broth Agar: 2% glucose, 0.17% yeast nitrogen base without amino acids (Difco), 0.5% ammonium sulphate, 1% casamino acids (Difco), 1.5% adenine sulphate, pH5.8 with acetic acid, 2% bacto agar (Difco). Autoclaved and stored at room temperature.

2xYT Broth: 1.6% Tryptone (Difco), 1% NaCl, 1% yeast extract (Difco), pH7.6 with NaOH. Autoclaved and stored at room temperature.
Superbroth: 3.5% Tryptone (Difco), 0.5% NaCl, 2% yeast extract (Difco), pH 7.5 with NaOH. Autoclaved and stored at room temperature.

Note: All media and most heat insensitive solutions were autoclaved using the following conditions: 120°C, 15lbs/in² for 15min.

2.3. *C.elegans*, yeast and bacterial strains

*C.elegans*: Wild-type Bristol N2 stock (Brenner, 1974).
Obtained from the *Caenorhabditis* Genetics Center (University of Minnesota, St Paul, Minnesota, USA)

Yeast: AB1380 (*MATα*, *ura3*, *trp1*, *ade2-1*, *can1-100*, *lys2-1*, *his5*).
Host strain for YAC clones. Gift from J.Sulston and A.Coulson.

*E.coli* strains:

<table>
<thead>
<tr>
<th>Name</th>
<th>Genotype</th>
</tr>
</thead>
<tbody>
<tr>
<td>ED8767</td>
<td>recA56, supE, <em>supF</em>, <em>hsdS</em>, <em>metE</em> Host cells for cosmid clones. Gift from J.Sulston and A.Coulson.</td>
</tr>
<tr>
<td>LE392</td>
<td>e14′(<em>McrA</em>), <em>hsdR514</em>, <em>supE44</em>, <em>supF58</em>, <em>lacY1</em>, <em>galK2</em>, <em>galT22</em>, <em>metB1</em>, <em>trpR55</em> Obtained from Stratagene Ltd. (Cambridge, UK).</td>
</tr>
<tr>
<td>Mc1061</td>
<td><em>araD139</em>, Δ(<em>ara-leu</em>)7696, Δ(<em>lac</em>)174, <em>galU</em>, <em>galK</em>, <em>hsdR2</em>(<em>K</em>, mK+), <em>mcrB1</em>, <em>rpsL</em>(<em>Str</em>) Gift from J.Sulston and A.Coulson.</td>
</tr>
</tbody>
</table>
OP50 The strain used in the our laboratory is a variant of the uracil requiring OP50 strain (Brenner, 1974) which has been transformed with a plasmid containing the Tet' marker (I.L. Johnstone, pers. comm.).

Popout A popout strain of E.coli. The strain is a P1 lysogen and a lambda lysogen and was used for plasmid excision from lambda SHLX2 clones. Gift from J.Sulston and A.Coulson.

SURE e14 (McrA'), Δ(mcrCB- hsdSMR- mrr)171, endA1, supE44, thi-1, gyrA96, relA1, lac, recB, recJ, sbcC, umuC::Tn5 (Kan'), uvrC, [F', proAB, lacF'ZΔM15, Tn10 (Tet')]. Obtained from Stratagene Ltd.

XL1-Blue recA1, endA1, gyrA96, thi-1, hsdR17, supE44, relA1, lac [F', proAB, lacF'ZΔM15, Tn10 (Tet')] Obtained from Stratagene Ltd.

2.4. Vectors and clones

2.4.1. Plasmid vectors

pBluescript II KS+:

pBluescript II KSc-:

pBluescript II SK-:

pPD21.28:

pPD95.03: lacZ fusion vector (Fire et al., 1990). Gift from A.Fire.

lacZ fusion vector (A.Fire, pers. comm.). Gift from A.Fire.

2.4.2. Phage

VCSM13: Derived from M13 K07 mutant. Obtained From Stratagene Ltd.
lambda EMBL4 C.elegans Bristol N2 Genomic DNA library. Gift from I.Johnstone.
2.4.3. Clones

pCe7: The cloned insert of pCe7 is a 7kb BamHI rDNA fragment containing a single repeat of the 18s and 26s ribosomal DNA sequences from *C. elegans* (Files and Hirsh, 1981). Gift from M.Krause.

pCEIF: pCEIF contains a 1615bp cDNA insert which encodes the *C. elegans* homologue of the eukaryotic initiation factor 4A (Roussell and Bennett, 1992). Gift from D.Roussell and K.Bennett.

pRF4: Contains the dominant rol-6(su1006) *C. elegans* cuticular collagen mutation (Mello et al., 1991). Gift from J.Kramer.

cDNA Clones: cm12b6, cm14e3, cm04d10 and cm01a5 (Waterston et al., 1992). Gift from J.Sulston and A.Coulson.


YAC Clones: Y69A3, Y55B10, Y43B9, Y7E11, Y50D4 and Y48G5. Gift from J.Sulston and A.Coulson.

2.5. Culture maintenance

2.5.1. *C. elegans* culture

Bristol N2 wild-type and transgenic *C. elegans* strains were maintained at 20°C on NGM agar plates with *E. coli* OP50 as a food source according to standard methods (Sulston and Hodgkin, 1988). For large scale growth of *C. elegans*, 500ml liquid
cultures were used according to standard methods (Sulston and Hodgkin, 1988). For mixed stage populations of *C. elegans*, worms were grown in 500ml liquid culture at 20°C with vigorous shaking for 3-4 days according to standard protocols (Sulston and Hodgkin, 1988). The culture was monitored at intervals by removing 1ml samples and examining the worms using a Zeiss Stemi 2000-C microscope (Carl Zeiss, Oberkochen, FRG).

2.5.2. Yeast culture

*Saccharomyces cerevisiae* strain AB1380 containing each of the YAC clones were maintained on YCD plates, or in 40ml of YPD media with vigorous shaking, at 30°C according to standard methods (Sherman *et al.*, 1986).

2.5.3. Bacterial culture

2.5.3.1. Bacterial culture on plates

All *E.coli* bacterial strains were grown at 37°C on L-agar plates, supplemented with the appropriate antibiotic at the appropriate concentration. The final concentrations of the antibiotics used were as follows:

i)  *E.coli* strains transformed with plasmids:  
   100μg/ml ampicillin or  
   12μg/ml tetracycline

ii) *E.coli* strains transformed with cosmids:  
    75μg/ml ampicillin or  
    30μg/ml kanamycin

2.5.3.2. Bacterial liquid culture

Unless otherwise stated, all liquid cultures were grown at 37°C with vigorous shaking. The media, antibiotics and antibiotic concentrations used in these cultures were as follows:
i) \textit{E.coli} strains transformed with plasmids:
L-broth supplemented with 100\(\mu\)g/ml ampicillin or 12\(\mu\)g/ml tetracycline.

ii) \textit{E.coli} strains transformed with cosmids:
2xYT supplemented with 120\(\mu\)g/ml ampicillin or 75\(\mu\)g/ml kanamycin.

2.6. DNA preparation

2.6.1. Plasmid DNA mini preparation

Magic and Wizard Miniprep kits (Promega Corporation, Southampton, UK), or QIAprep-spin kit (Qiagen Inc., via Hybaid Ltd., Teddington, Middlesex, UK), were used for plasmid DNA mini preparations from an appropriate volume of overnight culture of \textit{E.coli} cells, according to the manufacturer’s protocols supplied with these kits. Alternatively, plasmid DNA was isolated from 1.5ml of overnight cultures of \textit{E.coli} cells by alkaline extraction (Birnboim and Doly, 1979), organic extraction and ethanol precipitation according to standard protocols (Sambrook \textit{et al.}, 1989). The plasmid DNA was resuspended in an appropriate volume of TE or sterile distilled water, usually 50\(\mu\)l.

2.6.2. ssDNA template preparation for sequencing

i) 1ml of overnight culture of \textit{E.coli} cells containing the appropriate plasmid was used to inoculate 30ml of superbroth and grown at 37°C for 75min in a shaking incubator.

ii) 30\(\mu\)l of VCSM13 helper phage with a titre of approximately 1x10\(^{11}\) pfu/ml (Stratagene Ltd.) were added to the culture and incubated in the same conditions for a further 7 hours.

iii) The culture was centrifuged at 10,000 rpm for 10min using a Beckman J2.HS centrifuge and JA-17 rotor (Beckman Instruments Inc., Palo Alto, California, USA).

iv) Phage particles containing the ssDNA template were precipitated using 0.25 volumes of 20\% PEG-8000, 3.5M ammonium acetate at room temperature for 15min.
v) The precipitated phage particles were centrifuged 10,000 rpm for 10min as above.

vi) The phage particles were resuspended in 1ml of TE and reprecipitated as above.

vii) The pelleted phage particles were resuspended in 700μl of TE, phenol extracted (3 times), Phenol / chloroform extracted (1 time), chloroform extracted (1 time), ethanol precipitated, washed and resuspended in 15μl of TE using standard protocols (Sambrook et al., 1989).

2.6.3. Plasmid DNA medium and large scale preparations

Qiagen Plasmid Midi and Plasmid Maxi DNA Preparation kits (Qiagen Inc.) were used for medium and large scale plasmid DNA preparations respectively, from an appropriate volume of overnight culture of E.coli cells according to the manufacturer’s protocols.

2.6.4. Large scale cosmid DNA preparations

Large scale cosmid DNA preparations were made from 200-250ml cultures of E.coli cells containing the desired cosmid. The cultures were grown in two phases; a 20ml overnight culture was prepared and 10ml of this used to inoculate 200-250ml of fresh media which was grown to late log phase. Cosmid DNA was extracted using the alkaline hydrolysis method (Birnboim and Doly, 1979; Sambrook et al., 1989) with minor modifications communicated by Alan Coulson. The modifications are listed below:

i) Solution I was prepared with Tris-HCl pH7.4 and 8ml of this solution was used.

ii) No lysozyme step was used.

iii) 8ml of Solution II was used.

iv) Solution III was replaced with 3M Sodium Acetate pH4.8, 6ml were used and the mixture was left on ice for 30min.

v) Centrifugation of the precipitate generated by addition of 3M Sodium Acetate pH4.8 was performed at 12,000 rpm for 10min
The isopropanol precipitation and phenol/chloroform steps were omitted.

The precipitation step using 2 volumes of ethanol was performed at room temperature for 5 min, and no salt was added.

The cosmid DNA pellet was resuspended in 2.5 ml of 10 mM Tris-HCl pH 7.4, 10 mM Na₂EDTA overnight at 4°C.

Solution I: 50 mM glucose, 25 mM Tris-HCl (pH 8.0), 10 mM Na₂EDTA, 2 mg/ml lysozyme added fresh prior to use.

Solution II: 0.2 M NaOH, 1% SDS.

Solution III: 3 M sodium acetate (pH 4.8).

The cosmid DNA was purified using continuous caesium chloride / ethidium bromide density gradient centrifugation, isobutanol extraction and ethanol precipitation according to standard protocols (Sambrook et al., 1989).

### 2.6.5. Lambda phage DNA preparations

Lambda phage DNA was isolated and purified using the Promega Magic Lambda Prep kit (Promega Corporation) according to the manufacturer's protocol and using *E. coli* LE392 as the host strain. The *E. coli* LE392 host cells were cultured according to the liquid culture method of the protocol.

### 2.6.6. Yeast genomic DNA preparation

40 ml of yeast containing the appropriate yeast artificial chromosome (YAC) clones were grown in YPD media for 48 hours at 30°C. High molecular weight DNA (Genomic DNA and YAC DNA) was extracted and purified according to the yeast DNA miniprep protocol of Sherman *et al.* (1986) with one alteration: the zymolyase protoplasting step was replaced by treatment with 10 U of lyticase at 37°C for 4 hours.
2.7. RNA preparation

2.7.1. Isolation of mixed stage embryos

Worms were grown in 500ml liquid culture at 20°C with vigorous shaking until the E.coli OP50 food source was just clearing (approximately 4 days) according to standard protocols (Sulston and Hodgkin, 1988). The worms were purified from contaminating bacteria by sucrose floatation according to standard methods (Sulston and Hodgkin, 1988) with one exception: the worms were not left to settle overnight at 4°C prior to the sucrose floatation step. Eggs were isolated by treatment of the worms with hypochlorite according to standard protocols (Sulston and Hodgkin, 1988).

2.7.2. RNA extraction

RNA was extracted and purified from eggs (see Section 2.7.1) and mixed stage populations of C.elegans (see Section 2.5.1). RNA was extracted from eggs or mixed stage worms as follows:

i) Worms or eggs were ground in liquid nitrogen using a pestle and mortar.

ii) The ground material was added to a solution of 4M guanidinium thiocyanate, 0.13% sarkosyl, 33mM Tris-HCl pH 8.0, 0.5% β-mercaptoethanol and 6.7mM Na₂EDTA and homogenised thoroughly using a PTFE hand homogeniser.

iii) The solution was phenol (pH4.3) / chloroform extracted three times according to standard protocols (Sambrook et al., 1989).

iv) The RNA was precipitated with 0.3M sodium acetate and 0.75 volumes of ethanol

v) The pelleted RNA was resuspended in 10ml of the guanidinium thiocyanate solution described above, reprecipitated with 0.5 volumes of ethanol, centrifuged, washed and resuspended in 50-200µl of DEPC treated sterile distilled water according to standard methods (Sambrook et al., 1989).

Poly(A)⁺ selected RNA was obtained using the PolyATtract kit (Promega Corporation) according to the manufacturer’s protocol.
2.8. Standard molecular biology techniques

2.8.1. Restriction endonuclease digests

Single enzyme digests were carried out at 37°C in 1x enzyme reaction buffer supplied with the restriction endonuclease. Double and triple digests were performed in a buffer which yielded at least 50% activity for each of the enzymes used. There were situations where the enzymes used required incompatible reaction conditions. In these cases heat inactivatable enzymes were used first, inactivated by heating at 65°C for 10 min, the DNA precipitated, washed with ethanol and resuspended in sterile distilled water containing 1x reaction buffer appropriate for the next restriction endonuclease to be used. In other cases, the restriction enzymes could not be heat inactivated and therefore this step was replaced with organic extraction. The quantity of enzymes used varied depending upon the source of the DNA and also upon its required usage, however a general rule was observed which states that 1 unit of enzyme digests 1 μg of lambda DNA in 1 hour in a 20μl volume at the appropriate temperature. For genomic DNA to be used for Southern analysis or DNA to be used for preparation of subclones, a 5x excess of enzyme was used to ensure complete digestion of the DNA.

2.8.2. Agarose Gel electrophoresis of RNA and DNA

2.8.2.1. Agarose gels for electrophoresis of DNA samples

Restriction endonuclease digested DNA and PCR amplified DNA samples were electrophoresed through agarose gels. Gels ranging from 0.6% to 2% agarose were prepared by dissolving the appropriate amount of agarose in 1x TBE and adding 0.2μg/ml ethidium bromide. These gels were run in 1x TBE using gel electrophoresis tanks from GibcoBRL or Pharmacia (Milton Keynes, UK). The gels were generally run at a voltage equivalent to 3V/cm. All DNA samples were loaded into the wells of these gels using DNA sample buffer (see below). The electrophoresed DNA was visualised using a UV transilluminator (Appligene, Durham, UK) and a record of the gel made using a Polaroid camera or a digital imaging system (Appligene). Each gel contained at
least one lane of DNA size standards, either 1 kb ladder or lambda/HindIII (both from GibcoBRL)

**Fragment Sizes (in kilobases)**

<table>
<thead>
<tr>
<th></th>
<th>12.216</th>
<th>11.198</th>
<th>10.180</th>
<th>9.162</th>
<th>8.144</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 kb ladder</td>
<td>7.126</td>
<td>6.108</td>
<td>5.090</td>
<td>4.072</td>
<td>3.054</td>
</tr>
<tr>
<td></td>
<td>2.036</td>
<td>1.636</td>
<td>1.018</td>
<td>0.517</td>
<td>0.506</td>
</tr>
<tr>
<td></td>
<td>0.396</td>
<td>0.344</td>
<td>0.298</td>
<td>0.220</td>
<td>0.201</td>
</tr>
<tr>
<td></td>
<td>0.154</td>
<td>0.134</td>
<td>0.075</td>
<td></td>
<td></td>
</tr>
<tr>
<td>lambda / HindIII</td>
<td>23.130</td>
<td>9.416</td>
<td>6.557</td>
<td>4.361</td>
<td>2.322</td>
</tr>
<tr>
<td></td>
<td>2.027</td>
<td>0.564</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**10x DNA Sample buffer**

- 50% glycerol
- 0.5% bromophenol blue
- 0.5% xylene cyanol
- in 1X TBE

**2.8.2.2. Agarose gels for electrophoresis of RNA samples**

Denaturing 1% agarose gels were prepared in 1x MOPS running buffer and 2% formaldehyde (final concentration) according to the Quick Formaldehyde RNA Gel protocol of the ZAP-cDNA kit (Stratagene Ltd.). These gels were run in 1x MOPS running buffer using gel electrophoresis tanks from GibcoBRL. According to the Stratagene Quick Formaldehyde RNA Gel protocol, RNA samples (final volume 2μl) were mixed with 10μl of RNA sample buffer (see below) and boiled for 2min prior to loading. For total RNA samples, ethidium bromide was included in the sample buffer. For poly(A)+ selected RNA samples, the ethidium bromide was omitted. Each gel contained one lane of RNA size standards (GibcoBRL). The electrophoresed RNA was visualised using a UV transilluminator (Appligene) and a record of the gel made using either a Polaroid camera or an digital imaging system (Appligene).

**Fragment Sizes (in kilobases)**

<table>
<thead>
<tr>
<th>0.24-9.5 kb RNA Ladder</th>
<th>9.49</th>
<th>7.46</th>
<th>4.40</th>
<th>2.37</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1.35</td>
<td>0.24</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
RNA sample buffer

48% formamide
1x MOPS running buffer
6.4% formaldehyde
5.3% glycerol
5.3% of a saturated bromophenol blue solution
666μg/ml ethidium bromide (optional)

2.8.3. Gel purification

Gel purification of DNA fragments generated from restriction endonuclease digests or PCR reactions were performed in two ways. All fragments were electrophoresed through agarose gels run in 1x TBE. The appropriate fragments were excised using a scalpel blade and the DNA either extracted from the gel using the QIAEX DNA or QIAquick Gel Extraction kits (both from QIAGEN Inc.) according to the manufacturer's protocols.

2.8.4. Transformation of E.coli

Transformations were performed using E.coli XL1-blue supercompetent cells and 25-100ng of ligated DNA according to the manufacturers protocol with the following exception: Only 20μl of supercompetent cells were used for the transformations and the volumes of all solutions required for the transformation were adjusted accordingly. Alternatively, competent E.coli cells were prepared by treatment with 50mM CaCl₂ and transformed with plasmid DNA according to standard protocols (Sambrook et al., 1989).

The transformed cells were selected using the antibiotic resistance marker carried by the plasmid (Amp'). Where applicable, blue / white colony selection was also used to select transformants with plasmids containing inserts. This selection makes use of α-complementation of the lac2ΔM15 mutation in appropriate E.coli strains. In these instances, the agar media was also supplemented with 0.004% X-gal and 34μM IPTG (final concentrations).
2.8.5. Ligation reactions

Ligation reactions were performed in a 10μl or 15μl volume using 1x T4 DNA ligase buffer (Promega Corporation) and 1-2 units of T4 DNA ligase (Promega Corporation). Generally, 250ng of vector DNA were used in each ligation. The amount of insert DNA used varied with the molar concentration of free ends in the reaction. Ligations were performed overnight either at 16°C or at room temperature.

2.8.6. Southern and Northern blots

Northern blots and Southern blots were prepared from RNA and DNA electrophoresed through the appropriate type of agarose gel (see Section 2.8.2). All Northern and Southern Blots were performed using Hybond N (Amersham International plc, Little Chalfont, Buckinghamshire, UK) according to the manufacturers protocols. Southern blots were performed for 24 - 48 hours, Northern blots for 24 hours. The Southern blots were rinsed in 2x SSC and either baked at 80°C for 2 hours or UV cross-linked. The 2x SSC wash step was omitted for Northern blots and all Northern blots were UV cross-linked. All UV cross-linking was performed using a D-1000 cross-linker (Ultraviolet Products Ltd., Cambridge, UK).

2.8.7. Probe preparation

2.8.7.1. Random priming

All radiolabelled dsDNA probes were prepared by random priming using 50-100ng of DNA, (α-32P)dCTP (Dupont UK Ltd., Stevenage, UK) and the Prime-It or Prime-it II Kits (Stratagene Ltd.) according to the manufacturer’s protocols. All cloned cDNA probes were gel purified from their vectors prior to labelling. All PCR amplified probes were gel purified from other amplified bands prior to labelling.
2.8.7.2. 5' end labelling

Oligonucleotides (200ng) were end labelled using (γ-\(^{32}\)P)ATP (Dupont UK Ltd), 10U of polynucleotide kinase (Promega Corporation) and 1x polynucleotide kinase buffer supplied with the enzyme in a final reaction volume of 10μl. The kinase reaction was allowed to proceed at 37°C for 45min.

2.8.7.3. Riboprobes

Antisense riboprobes were prepared using the T3 or T7 RNA polymerases, (γ-\(^{32}\)P)rUTP (Dupont) and the RNA Transcription Kit (Stratagene Ltd.) according to the manufacturer’s protocols but with the addition of RNASin Ribonuclease inhibitor (Promega Corporation).

2.8.7.4. Probe purification

All probes were purified using NucTrap Probe Purification Columns (Stratagene Ltd.) according to the manufacturer’s protocols.

2.8.8. Hybridisation

All hybridisations were performed in Hybaid (Hybaid Ltd., Teddington, Middlesex, UK) hybridisation bottles using Bachofen (Reutlingen, FRG) hybridisation ovens. All Northern and Southern blots were stripped using a boiling solution of 0.1% SDS after exposure to medical X-ray film.

2.8.8.1. Southern hybridisations

All Southern hybridisations were performed according to the Amersham Hybond N protocols (Amersham International plc) with the following alterations:
The solution used for prehybridisation and hybridisation was altered to:

- 6x SSC
- 5x Denhardt's Solution
- 0.5% SDS
- 20μg/ml Herring testes DNA (Denatured by heating at 100°C for 10min prior to addition to the solution)

Prehybridisations were performed for at least 3 hours using 25ml of the above solution

Prior to addition of the radiolabelled probe, sufficient prehybridisation solution was removed such that the remainder barely covered the Southern blots.

All radiolabelled, random primed dsDNA probes were denatured by heating at 100°C for 5min prior to addition to the hybridisation bottles. All Southern hybridisations were performed for at least 16 hours at 65°C unless 5' end-labelled oligonucleotide probes were being used, in which case the hybridisations were performed for at least 16 hours at 55°C.

### 2.8.8.2. Northern hybridisations

The Northern blots were rolled up and inserted, RNA side up, into the hybridisation bottles. The Northern blots were prehybridised and hybridised in the same solution. The prehybridisation / hybridisation solution comprised, 50% formamide, 5x SSPE, 5x Denhardts, 0.5% SDS and 100μg/ml herring testes DNA. Prehybridisations were performed for at least 4 hours and hybridisations were performed for at least 16 hours. Both prehybridisation and hybridisation were performed at 42°C. Prior to addition of the radiolabelled probe, sufficient prehybridisation solution was removed such that the remainder barely covered the Northern blots. Random primed cDNA probes were denatured at 100°C for 10min and riboprobes were heated at 65°C for 5min prior to addition to the hybridisation solution.
2.8.8.3. Southern hybridisation washes

All hybridised Southern blots were washed once at low stringency in at least 1 litre of 2x SSC / 0.1% SDS solution at room temperature for 30 min. All Southern blots probed with random primed DNA probes were washed with two high stringency washes using at least 1 litre of 0.1x SSC / 0.1% SDS solution at 65°C for 15 min. For Southern blots probed with 5' end-labelled oligonucleotide probes, the two high stringency washes were replaced with a single medium stringency wash performed using at least 1 litre of 0.5x SSC / 0.1% SDS solution at 55°C for 15 min.

2.8.8.4. Northern hybridisation washes

Both total RNA Northern blots probed with random primed cDNA probes and poly(A)+ RNA Northern blots probed with riboprobes were initially washed using at least 1 litre of 2x SSPE / 0.1% SDS solution at room temperature for 20 min. All Northern filters were subsequently washed at medium stringency using at least 1 litre of 1x SSPE / 0.1% SDS solution at 65°C for 15 min. A high stringency wash was also performed where required, using at least 1 litre of 0.1x SSPE / 0.1% SDS solution at 65°C for 10 min.

2.8.8.5. Autoradiography

Probed Southern and Northern blots were sealed in bags and exposed to sheets of Agfa Curix RP1 medical X-ray film (via H.A. West, Clydebank, Dumbartonshire, UK) in autoradiography cassettes with intensifying screens. Exposures were carried out at 70°C until satisfactory images were produced. The autoradiography film was developed using an M35-M X-omat processor (Kodak Ltd., Hemel Hempstead, Hertfordshire, UK).
2.8.8.6. Colony hybridisation

Colony hybridisation was performed using replica plates. Each colony to be screened was streaked onto two L-agar plates, both supplemented with the appropriate antibiotic and one overlaid with an 88mm Hybond N (Amersham International plc) Nylcor filter. A template was used to streak each colony onto the surface of the two plates in an ordered grid. Thus, each colony was first streaked onto the master plate at a position defined by the template. The colony was subsequently streaked onto the Hybond N filter (Amersham International plc) of the second plate at the equivalent position. The master and replica plates were grown overnight at 37°C. The bacterial colonies on the Nylcor filter were lysed and the DNA was denatured and neutralised according to standard protocols for colony hybridisation (Sambrook et al., 1989). The DNA was cross-linked to the filters by baking for 2 hours at 80°C. The resulting filters were hybridised according to the conditions described for Southern hybridisation (Section 2.8.8.1) using random-primed cDNA probes (Section 2.8.7.1) and were washed twice at high stringency (Section 2.8.8.3) prior to autoradiography (Section 2.8.8.5).

2.9. Screening a lambda EMBL4 C.elegans Bristol N2 genomic DNA library

E.coli LE392 was used as a host strain for phage of the lambda EMBL4 library. The cells were cultured in L-broth supplemented with 0.2% glucose and 10mM MgSO₄ at 37°C overnight. The cells were centrifuged at 3,000 rpm using a Beckman J2-HS centrifuge and JA-17 rotor (Beckman Instruments Inc.). The pelleted cells were resuspended in 10mM MgSO₄ to generate a suspension with an OD₆₀₀ of 2.

2.9.1. Determining the titre of the lambda library

The titre of the lambda EMBL4 library (prepared by Dr I.L.Johnstone) was determined using standard methods (Sambrook et al., 1989) with the following exceptions:

1) 100μl of E.coli LE392 host cells were transfected with 10μl of each serial dilution (1:1-1:10,000) of lysate from the amplified library.
ii) The infected cells were plated onto BBL bottom agar, in 9cm petri dishes, using 3ml of BBL top agarose.

2.9.2. Plating and screening of the lambda library

Approximately 90,000 lambda phage from the lambda EMBL4 library (15,000/plate) were plated and screened according to standard methods (Sambrook et al., 1989), with the following exceptions:

i) Initially, 50µl of phage library lysate (approximately 15,000 phage) were mixed with 100µl of E.coli LE392 host cells

ii) The infected cells were plated onto BBL bottom agar, in 9cm petri dishes, using 3ml of BBL top agarose.

iii) Hybond N nylon filters (Amersham International plc) were used for the transfer of phage DNA.

iv) After denaturation and neutralisation, the filters were baked at 80°C for 2 hours to fix the DNA to the filters.

v) The immobilised DNA was hybridised with a ³²P-labelled cm12b6 cDNA probe (radiolabelled by random priming, Section 2.8.7.1) using the conditions described in Section 2.8.8.1. The filters were washed twice at high stringency (Section 2.8.8.3) prior to autoradiography (Section 2.8.8.5).

2.10. Excision of cDNA clones from lambda SHLX2

The pRATII plasmids containing the cDNA clones were excised from lambda SHLX2 phage according to the instructions supplied with the phage clones. Lysates from the lambda SHLX2 cDNA clones cm12b6, cm14e3, cm04d10 and cm01a5 (Waterston et al., 1992) were obtained from J.Sulston and A.Coulson. E.coli MC1061 cells were grown at 37°C with vigorous shaking until late log phase. The cells were pelleted and resuspended in 10mM MgSO₄ to generate a suspension with an OD₆₀₀ of 1 (as described in Section 2.9). 100µl of the cell suspension were transfected with 1µl of 10-fold serial dilutions of lysate (1:1-1:1,000) from each of the four cDNA clones and
incubated at 37°C for 15min. The infected cells were plated using standard methods (Sambrook et al., 1989), except that 3ml of BBL top agarose were used to plate the infected cells onto BBL bottom agar, in 9cm petri dishes. The plates were incubated at 37°C overnight. Agarose plugs from single well isolated plaques for each of the cDNA clones were obtained, eluted in 1ml of SM buffer and stored at 4°C with 10μl of chloroform according to standard methods (Sambrook et al., 1989). 10μl of each resulting plaque lysate were added to 100μl of an overnight culture of a popout strain of E.coli (P1 and lambda lysogen) and 10μl of VCSM13 helper phage (titre: 6.1x10^{11} pfu/ml), incubated at room temperature for 75min, added to 2ml of 2xYT supplemented with ampicillin (90μg/ml) and incubated at 37°C overnight. The overnight culture was centrifuged and 5μl of the resulting supernatant mixed with 100μl of an overnight culture of E.coli XL1-blue, incubated at room temperature for 75min, streaked onto L-agar plates supplemented with ampicillin (90μg/ml) and tetracycline (15μg/ml) and grown at 37°C overnight. Large isolated colonies were selected for further analysis. The cDNA inserts were subsequently subcloned as Apal/Sacl fragments into pBluescript II KS+ for further analysis.

2.11. Nucleotide sequencing

All sequencing was performed using the dideoxynucleotide chain termination method (Sanger et al., 1977) using the Sequenase version 2.0 kit or the Sequenase version 2.0 Quick Denature kits and (α-^{35}S)dATP (Dupont UK Ltd.) according to the manufacturer’s protocols (Amersham International plc). The sequencing reactions were performed using single or double stranded DNA templates and either custom made sequencing primers or primers that anneal to sequences flanking the pBluescript polylinker, into which the cDNA and genomic clones were inserted. The sequencing reactions were electrophoresed at 70 watts through a denaturing 6% acrylamide gel (see below) with 1xTBE using Model S2 (GibcoBRL) or Base Ace sequencing gel tanks (Stratagene Ltd.). The sequencing gels were fixed in 10% methanol, 5% glacial acetic acid at room temperature for 30min. The gels were dried onto 3MM paper at 80°C for 2 hours using a Bio-Rad Model 583 gel drier (Bio-Rad Laboratories Ltd., Hemel Hempstead, Hertfordshire, UK). The dried gels were exposed directly to sheets of Agfa
Curix RP1 medical X-ray film (via H.A.West) in autoradiography cassettes with intensifying screens, exposures were carried out at room temperature until satisfactory images were produced. The autoradiography film was developed using an M35-M X-omat processor (Kodak Ltd.).

Sequencing gel:

- 50% UREA
- 6% acrylamide (19:1 Acrylamide/Bis-acrylamide, Anachem, Luton, Bedfordshire, UK)
- 1xTBE (replace with 1xTTE for glycerol tolerant gels)
- 0.1% ammonium persulphate
- (polymerisation of acrylamide effected by addition of TEMED to a final concentration of 0.01%).

2.11.1. Preparation of custom made primers

Custom oligonucleotide primers for nucleotide sequencing and Polymerase Chain Reaction (PCR) were obtained from two sources. They were either synthesised by myself using a PCR Mate DNA synthesiser (Applied Biosystems, Warrington, Cheshire, UK) or were supplied ready to use by Strathclyde University (Glasgow, UK).

2.11.2. Deprotection of synthesised oligonucleotides

The protective group of primers synthesised using the Applied Biosystems PCR Mate DNA synthesiser must be removed prior to use. The primers were deprotected and purified using the following protocol:

i) The beads supporting the synthesised oligonucleotides were transferred from the column into a 1.5ml screw cap eppendorf; 1ml of ammonium hydroxide (30% w/v) was added and the tube was incubated at room temperature for 2 hours with occasional vortexing.

ii) The eppendorfs were centrifuged at 15,000 rpm for 5 minute in a microfuge (Heraeus Equipment Ltd., Brentwood, Essex, UK), the supernatant was transferred to a new screw cap eppendorf, ammonium hydroxide (30% w/v)
was added to the top and the tube was incubated overnight at 55°C.

iii) The solution was transferred to a 12ml polypropylene centrifuge tube and precipitated using 0.1 volumes of 5M ammonium acetate and 3 volumes of ethanol for 1 hour at -20°C.

iv) The precipitated DNA was pelleted by centrifuging at 6,000 rpm for 10min using a J2-HS centrifuge and JA-17 rotor (Beckman Instruments plc)

v) The oligonucleotides were resuspended in 500µl of TE, precipitated with 0.1 volumes of 3M sodium acetate pH5.2 and 3 volumes of ethanol for 1 hour at 20°C and centrifuged as above.

vi) The oligonucleotides were washed in 70% ethanol, dried and resuspended in 200-500µl of TE.

2.11.3. Nucleotide sequence of the primers used for sequencing

2.11.3.1. Primers for sequencing the 5' rapid amplification of cDNA ends (RACE) products

All the cloned 5' RACE products were sequenced using primers flanking the pBluescript II KS+ vector polylinker. The primers were all obtained from Stratagene Ltd. and are listed below:

<table>
<thead>
<tr>
<th>Name</th>
<th>Primer sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>T3 17mer</td>
<td>5'-GAAATCACCTCCCAATTATTA-3'</td>
</tr>
<tr>
<td>T7 17mer</td>
<td>5'-AATACGACTGACATGATAG-3'</td>
</tr>
<tr>
<td>sk 17mer</td>
<td>5'-CTAGGTGAACCAAGATCT-3'</td>
</tr>
<tr>
<td>ks 17mer</td>
<td>5'-CGAGGTGACGCTATCG-3'</td>
</tr>
</tbody>
</table>

2.11.3.2. Primers for sequencing the cDNA and genomic clones of crp-3, crp-4, crp-5 and crp-6

The nucleotide sequences of the cDNA and genomic clones were obtained using a combination of primers flanking the pBluescript vector polylinker and custom made primers. Most of the genomic clone nucleotide sequences were obtained using custom made primers, however some sequence data were obtained using the T7 17mer whose
nucleotide sequence is given above. Most of the nucleotide sequence of the cDNA clones was also obtained using custom primers however some sequence data were also obtained using the T3 17mer (listed above) and the M13 (-40 Forward) primer supplied with the Sequenase kits (Amersham International plc). The nucleotide sequence of this primer is listed below:

<table>
<thead>
<tr>
<th>Name</th>
<th>Primer sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>M13 (-40 Forward)</td>
<td>5'-GTTTTCCCAGTCACGACGTTGTA-3'</td>
</tr>
</tbody>
</table>

The nucleotide sequences of the custom made primers used to obtain the majority of the nucleotide sequences for the cDNA and genomic clones of each of the four genes are listed below. The position of the primers with respect to the ATG initiation codons of each of the four genes is also given to help understand how the final sequence of each of the four genes (Chapter 4, Figures 4.11, 4.12, 4.13 and 4.14) was obtained.
### Table i. Custom primers for *cpr-3*

<table>
<thead>
<tr>
<th>Name</th>
<th>Primer (Read 5'-3')</th>
<th>Clones Sequenced With Primer</th>
<th>Sense of Primer</th>
<th>Position of 5' End and 3' End Bases of Each Primer With Respect to ATG Initiation Codon</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>5' End</td>
</tr>
<tr>
<td>cpr-3/0.5</td>
<td>AGTTTTCAGTTTTCCCTC</td>
<td>genomic</td>
<td>sense</td>
<td>-46</td>
</tr>
<tr>
<td>cpr-3/1</td>
<td>CGCGCGGAATCAATTG</td>
<td>cDNA and genomic</td>
<td>sense</td>
<td>+114</td>
</tr>
<tr>
<td>cpr-3/2</td>
<td>CACCGTACAAAACCAGCT</td>
<td>cDNA and genomic</td>
<td>sense</td>
<td>+168</td>
</tr>
<tr>
<td>cpr-3/3</td>
<td>TACGGATGTAAGGAGGC</td>
<td>cDNA and genomic</td>
<td>sense</td>
<td>+526</td>
</tr>
<tr>
<td>cpr-3/int/2.5</td>
<td>ACGAGAATTATTATCAGATG</td>
<td>genomic</td>
<td>sense</td>
<td>+806</td>
</tr>
<tr>
<td>cpr-3/4</td>
<td>AATCTACCAATTACGGAC</td>
<td>cDNA and genomic</td>
<td>sense</td>
<td>+1198</td>
</tr>
<tr>
<td>cpr-3/5</td>
<td>ACTATACTTCCGGAAGC</td>
<td>cDNA and genomic</td>
<td>sense</td>
<td>+1275</td>
</tr>
<tr>
<td>cpr-3/6</td>
<td>GAGTGCCTCAGATGATG</td>
<td>cDNA and genomic</td>
<td>sense</td>
<td>+1671</td>
</tr>
<tr>
<td>cpr-3/R0.5</td>
<td>GAAAACTGGAAAGAAGCG</td>
<td>genomic</td>
<td>antisense</td>
<td>-40</td>
</tr>
<tr>
<td>cpr-3/R1</td>
<td>GCGTCGACCTGGTTTGA</td>
<td>cDNA and genomic</td>
<td>antisense</td>
<td>+182</td>
</tr>
<tr>
<td>cpr-3/R2</td>
<td>TCAACGGTTTGTAGTTCCATT</td>
<td>cDNA and genomic</td>
<td>antisense</td>
<td>+481</td>
</tr>
<tr>
<td>cpr-3/int/R2.5</td>
<td>CGTGAACGTGACCTCC</td>
<td>genomic</td>
<td>antisense</td>
<td>+944</td>
</tr>
<tr>
<td>cpr-3/R3</td>
<td>CGGTGACGGATTTCGTA</td>
<td>cDNA and genomic</td>
<td>antisense</td>
<td>+1184</td>
</tr>
<tr>
<td>cpr-3/R4</td>
<td>CAATTCCAGCTACCAGG</td>
<td>cDNA and genomic</td>
<td>antisense</td>
<td>+1460</td>
</tr>
<tr>
<td>cpr-3/R5</td>
<td>CGTCAAGCACCATCAATG</td>
<td>genomic</td>
<td>antisense</td>
<td>+1596</td>
</tr>
<tr>
<td>cpr-5'FL</td>
<td>GTACAATTCGATAATCT</td>
<td>genomic</td>
<td>sense</td>
<td>see below</td>
</tr>
<tr>
<td>cpr-3'FL</td>
<td>GAACTTCACTGATAATAC</td>
<td>genomic</td>
<td>antisense</td>
<td>see below</td>
</tr>
<tr>
<td>Name</td>
<td>Primer (Read 5’-3’)</td>
<td>Clones Sequenced With Primer</td>
<td>Sense of Primer</td>
<td>Position of 5’ End and 3’ End Bases of Each Primer With Respect to ATG Initiation Codon</td>
</tr>
<tr>
<td>--------</td>
<td>----------------------------------------------------------</td>
<td>------------------------------</td>
<td>-----------------</td>
<td>------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>cpr-4/0.5</td>
<td>TTGCTCTATCTTGTGCTATTGCTCTT</td>
<td>genomic</td>
<td>sense</td>
<td>-33 -9</td>
</tr>
<tr>
<td>cpr-4/1</td>
<td>GCCCTTGGTGCTTGCTTAC CGCCGG</td>
<td>cDNA and genomic</td>
<td>sense</td>
<td>+71 +93</td>
</tr>
<tr>
<td>cpr-4/2</td>
<td>CTATCACCGAGTATGCTG ACTCAACGGCAA</td>
<td>cDNA and genomic</td>
<td>sense</td>
<td>+129 +157</td>
</tr>
<tr>
<td>cpr-4/3</td>
<td>TCTTGCTGCTCAAACTGCG</td>
<td>cDNA and genomic</td>
<td>sense</td>
<td>+470 +488</td>
</tr>
<tr>
<td>cpr-4/4</td>
<td>CACCGCTTTACGCGCCGTCG GAAAG</td>
<td>cDNA and genomic</td>
<td>sense</td>
<td>+739 +760</td>
</tr>
<tr>
<td>cpr-4/5</td>
<td>GGAACCAACGAGTGC GG</td>
<td>genomic</td>
<td>sense</td>
<td>+1003 +1017</td>
</tr>
<tr>
<td>cpr-4/R0.5</td>
<td>AAAAGAGCAGAAATAGCAGG</td>
<td>genomic</td>
<td>antisense</td>
<td>-7 -24</td>
</tr>
<tr>
<td>cpr-4/R1</td>
<td>GCGTCGACCTTCCAGAG AGATTGCTTTGA</td>
<td>cDNA and genomic</td>
<td>antisense</td>
<td>+170 +149</td>
</tr>
<tr>
<td>cpr-4/R2</td>
<td>TTGGAGCGAGCAAGAAAG AACATC</td>
<td>cDNA and genomic</td>
<td>antisense</td>
<td>+483 +461</td>
</tr>
<tr>
<td>cpr-4/R2.5</td>
<td>CCGTGCGAGAATCCGCTC</td>
<td>genomic</td>
<td>antisense</td>
<td>+558 +541</td>
</tr>
<tr>
<td>cpr-4/R3</td>
<td>ACTGTCGCGTCGGGAATG</td>
<td>cDNA and genomic</td>
<td>antisense</td>
<td>+804 +787</td>
</tr>
<tr>
<td>cpr-4/R4</td>
<td>TATTTATTATCATATATATATACAG</td>
<td>cDNA and genomic</td>
<td>antisense</td>
<td>+1116 +1091</td>
</tr>
<tr>
<td>cpr-4/5’FL</td>
<td>TTCGCCGAGTACAGAG</td>
<td>genomic</td>
<td>sense</td>
<td>see below</td>
</tr>
<tr>
<td>cpr-4/3’FL</td>
<td>ATATCTCTAACCTTATCG</td>
<td>genomic</td>
<td>antisense</td>
<td>see below</td>
</tr>
</tbody>
</table>
### Table iii. Custom primers for cpr-5

<table>
<thead>
<tr>
<th>Name</th>
<th>Primer (Read 5’-3’)</th>
<th>Clones Sequenced With Primer</th>
<th>Sense of Primer</th>
<th>Position of 5’ End and 3’ End Bases of Each Primer With Respect to ATG Initiation Codon</th>
</tr>
</thead>
<tbody>
<tr>
<td>cpr-5/1</td>
<td>GCTTTGGACCGCTGGAC ATCAAG</td>
<td>cdDNA and genomic</td>
<td>sense</td>
<td>+123 +145</td>
</tr>
<tr>
<td>cpr-5/2</td>
<td>GGAGAAGATCCAAGAAGCTGATGG</td>
<td>cdDNA and genomic</td>
<td>sense</td>
<td>+156 +181</td>
</tr>
<tr>
<td>cpr-5/3</td>
<td>ACCGGAATGTTCACGTGCG</td>
<td>cdDNA and genomic</td>
<td>sense</td>
<td>+433 +451</td>
</tr>
<tr>
<td>cpr-5/4</td>
<td>CGAGGACCTTACCAATACCC</td>
<td>cdDNA and genomic</td>
<td>sense</td>
<td>+1026 +1047</td>
</tr>
<tr>
<td>cpr-5/5</td>
<td>CAGGGAGATTTCCATGTG</td>
<td>genomic</td>
<td>sense</td>
<td>+1303 +1320</td>
</tr>
<tr>
<td>cpr-5/R0.5</td>
<td>GAATAGCGGAGAGCTTC</td>
<td>genomic</td>
<td>antisense</td>
<td>+22 +5</td>
</tr>
<tr>
<td>cpr-5/R1</td>
<td>GCAAATGTGCAAGTGGC TCGGAATAGCGTC</td>
<td>cdDNA and genomic</td>
<td>antisense</td>
<td>+260 +238</td>
</tr>
<tr>
<td>cpr-5/R2</td>
<td>GATTTAAACTACGGC</td>
<td>genomic</td>
<td>antisense</td>
<td>+563 +547</td>
</tr>
<tr>
<td>cpr-5/R3</td>
<td>TTCCACAGTCTCAGGCGATGGA</td>
<td>cdDNA and genomic</td>
<td>antisense</td>
<td>+774 +753</td>
</tr>
<tr>
<td>cpr-5/FL</td>
<td>TTAGGATTAGGGCTCATC</td>
<td>genomic</td>
<td>antisense</td>
<td>+1127 +1110</td>
</tr>
<tr>
<td>cpr-5/3'FL</td>
<td>GAAATCAAGATTTGGG</td>
<td>genomic</td>
<td>antisense</td>
<td>see below</td>
</tr>
</tbody>
</table>


<table>
<thead>
<tr>
<th>Name</th>
<th>Primer (Read 5’-3’)</th>
<th>Clones Sequenced With Primer</th>
<th>Sense of Primer</th>
<th>Position of 5’ End and 3’ End Bases of Each Primer With Respect to ATG Initiation Codon</th>
</tr>
</thead>
<tbody>
<tr>
<td>cpr-6/0.5</td>
<td>CAAGCGACGACAACCTTG C</td>
<td>genomic</td>
<td></td>
<td>5’ End -45  3’ End -29</td>
</tr>
<tr>
<td>cpr-6/1</td>
<td>GCATAGTGTTAGCAGCT TATTGCCG</td>
<td>cDNA and genomic</td>
<td>sense +75</td>
<td>+99</td>
</tr>
<tr>
<td>cpr-6/2</td>
<td>GGACAAATATCAGCAATC GTGAAATTGACTC</td>
<td>cDNA and genomic</td>
<td>sense +127</td>
<td>+203</td>
</tr>
<tr>
<td>cpr-6/2.5</td>
<td>TGGGGATGGATGGTGTG C</td>
<td>genomic</td>
<td>sense +368</td>
<td>+385</td>
</tr>
<tr>
<td>cpr-6/3</td>
<td>GATCCCGTGGGGCTTTC TGGG</td>
<td>cDNA and genomic</td>
<td>sense +724</td>
<td>+742</td>
</tr>
<tr>
<td>cpr-6/4</td>
<td>CCAAAATGTGAAAGGA A GTGCC</td>
<td>cDNA and genomic</td>
<td>sense +1077</td>
<td>+1098</td>
</tr>
<tr>
<td>cpr-6/5</td>
<td>AGTCTATGGTTACACCC G</td>
<td>cDNA and genomic</td>
<td>sense +1262</td>
<td>+1326</td>
</tr>
<tr>
<td>cpr-6/5.5</td>
<td>CGGAGGGAGGACAGCGC G</td>
<td>genomic</td>
<td>sense +1336</td>
<td>+1352</td>
</tr>
<tr>
<td>cpr-6/5.75</td>
<td>TCTGGGATTGTTGGAGG G</td>
<td>genomic</td>
<td>sense +1481</td>
<td>+1497</td>
</tr>
<tr>
<td>cpr-6/6</td>
<td>CCGCCCGCACCTCTACG G</td>
<td>genomic</td>
<td>sense +1689</td>
<td>+1705</td>
</tr>
<tr>
<td>cpr-6/R0.5</td>
<td>TTGCGCTGACGCGTCAG C</td>
<td>genomic</td>
<td>antisense 2</td>
<td>-19</td>
</tr>
<tr>
<td>cpr-6/R1</td>
<td>GCACTAGTCCGTCATTG T TTATTACG</td>
<td>cDNA antisense</td>
<td>+330</td>
<td>+266</td>
</tr>
<tr>
<td>cpr-6/R1.5</td>
<td>GCCCTGCTCTTCCTCC C</td>
<td>genomic</td>
<td>antisense +363</td>
<td>+347</td>
</tr>
<tr>
<td>cpr-6/R1.75</td>
<td>AGCTTCCTGGAAATTGC C</td>
<td>genomic</td>
<td>antisense +657</td>
<td>+641</td>
</tr>
<tr>
<td>cpr-6/R2</td>
<td>GTTACATCCGAATTCCAC AGC</td>
<td>cDNA and genomic</td>
<td>antisense +857</td>
<td>+838</td>
</tr>
<tr>
<td>cpr-6/R3</td>
<td>CCGTGAGTCATCAATTCT C</td>
<td>cDNA and genomic</td>
<td>antisense +1207</td>
<td>+1190</td>
</tr>
<tr>
<td>cpr-6/R4</td>
<td>ACACCTCATACTCCTCC C</td>
<td>cDNA and genomic</td>
<td>antisense +1471</td>
<td>+1455</td>
</tr>
<tr>
<td>cpr-6/R5</td>
<td>TAATCTAATTCATATTATA CATAAG</td>
<td>cDNA antisense</td>
<td>+1788</td>
<td>+1763</td>
</tr>
<tr>
<td>cpr-6/5 FL</td>
<td>ATCGGTGCATTCACTGA T A</td>
<td>genomic</td>
<td>sense see below</td>
<td>see below</td>
</tr>
<tr>
<td>cpr-6/3 FL</td>
<td>TTCTAAATATATAGATT G</td>
<td>genomic</td>
<td>antisense see below</td>
<td>see below</td>
</tr>
</tbody>
</table>

Notes: The CPR-3/R1, CPR-4/R1, CPR-5/R1 and CPR-6/R1 primers all have restriction endonuclease sites engineered into their 5’ ends (shown in bold) because these primers were also used for the 5’ RACE experiments (see below). These sites did not therefore perform any function when these primers were used for sequencing.

The ~5’ FL and ~3’ FL primers for each of the four genes were used to sequence the second strand DNA of the 5’ and 3’ flanks of each of the four genes. The nucleotide
sequences of these primers were derived from regions of readable sequence outwith the 5' and 3' flank sequence of each gene obtained from first strand DNA sequencing. This approach was used in order to obtain more nucleotide sequence confirmed in both strands for the 5' and 3' flanks of each gene without having to prepare more custom made primers. Since these primers were derived from readable sequence that was not adjacent to that obtained by first strand DNA sequencing of the 5' and 3' flanks of each gene, the positions of these primers with respect to the ATG initiation codon are not known accurately and are therefore not included in the above tables. However, the approximate positions of the ~5'FL and ~3'FL primers are upstream of the 5' flank sequences and downstream of the 3' flank sequences given for each gene, respectively (shown in Chapter 4, Figures 4.11, 4.12, 4.13 and 4.14). The primer cpr-5/5'FL is the notable exception to this rule. First strand sequencing of the 5' flank of cpr-5 revealed much repetitive sequence which negated the design of a specific primer for obtaining second strand DNA sequence. Accordingly, this primer was designed to unique sequence downstream of the region of repetitive sequence and within the region of DNA sequence obtained by first strand DNA sequencing. Therefore, the position of this primer with respect to the ATG initiation codon is known and is included in the appropriate table above.

2.11.3.3. Primers for sequencing the cpr-3::lacZ, cpr-4::lacZ, cpr-5::lacZ and cpr-6::lacZ constructs

During the construction of the lacZ gene fusions, a number of primers were used to check the sequence of the PCR amplified linkers and to check the final constructs. The names and sequences of these primers are listed below.

i) Primers used to sequence the PCR amplified linker regions

<table>
<thead>
<tr>
<th>Name</th>
<th>Primer Sequence</th>
<th>Clones Sequenced</th>
</tr>
</thead>
<tbody>
<tr>
<td>T3 and T7 17mers</td>
<td>sequence given above</td>
<td>cpr-3, cpr-4 and cpr-6 PCR amplified linkers in pBluescript II SK-</td>
</tr>
<tr>
<td>cpr-5 Primer 'X'*</td>
<td>5'-TTAGGTATAGGCTCATC-3'</td>
<td>cpr-5::lacZ</td>
</tr>
</tbody>
</table>
SYNINT** 5'-GTTCTATGTTATGT  cpr-5::lacZ
TAGTATCATTCAGAAACAC-3'

*This primer is also known as cpr-5/5'FL and is listed in the appropriate table above
**This primer anneals to the synthetic intron of the pPD21.28 lacZ fusion vector (see Chapter 5, Figure 5.21) and was used to obtain antisense sequence for the cpr-5 PCR amplified linker.

ii) Custom made primers used to check cpr-3::lacZ, cpr-4::lacZ, cpr-5::lacZ and cpr-6::lacZ fusion constructs

<table>
<thead>
<tr>
<th>Name</th>
<th>Primer Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>NUCLOC</td>
<td>5'-TCACCCACCCGTACCTACGC-3'</td>
</tr>
<tr>
<td>M13 Reverse PCR Primer</td>
<td>5'-AACAGCTATGACCATGATTA-3'</td>
</tr>
<tr>
<td>SYNINT</td>
<td>sequence given above</td>
</tr>
</tbody>
</table>

Note: The orientation and relative positions of the NUCLOC, SYNINT and M13 Reverse PCR primers are shown schematically in Chapter 5, Figure 5.21.

2.12. 5' Rapid amplification of cDNA ends (5' RACE)

5' RACE was performed with 2μg of total RNA isolated from mixed stage N2 Bristol C.elegans and 1pg of GibcoBRL control RNA using the GibcoBRL 5' RACE kit (GibcoBRL) according to the manufacturer's protocol. The control RNA is an 891bp in-vitro transcribed region of the Chloramphenicol Acetyl Transferase gene, engineered to include a poly(A) tail. In brief, first strand cDNA was prepared from the total RNA (or from the control RNA) using pooled primers specific for each of the four genes (or the GSP1 control primer). The resulting cDNA was dC-tailed using Terminal Deoxynucleotidyl Transferase. The dC-tailed cDNA was subsequently amplified by nested PCR. The first round of the nested PCR used pooled primers specific to each of the four genes and the Anchor Primer supplied with the kit, which annealed to the dC-tailed region of the cDNAs. The cDNA generated from the GibcoBRL 5' RACE control
RNA was amplified in a separate reaction using the GSP2 control primer, the Universal
Primer and the same cycle profiles as for the four genes.

Four independent PCR reactions were performed on the amplified products from
the first round of nested PCR. Each reaction contained the Universal Primer, which
annealed to the 5' ends of the amplified products generated by the Anchor Primer in the
first round of PCR, and a primer specific to one of the four genes. The products of the
second round of nested PCR were electrophoresed through a 1.5% agarose gel and the
predominant products from each reaction were gel purified. The gel purified products
were cloned into pBluescript II SK- by means of SpeI or SalI restriction endonuclease
sites engineered into the 5' ends of the universal and gene specific primers used in the
final PCR amplification step. Thus the 5' RACE products of cpr-3 and cpr-4 were
cloned using SalI sites and those of cpr-5 and cpr-6 were cloned using SpeI sites. The
Anchor Primer, Universal Primer, control and gene specific primers used in the cDNA
synthesis and amplification steps are listed below:

<table>
<thead>
<tr>
<th>Name</th>
<th>Primers For cDNA Synthesis (GSP1)</th>
<th>Primers For 1st Round PCR (GSP2)</th>
<th>Primers For 2nd Round PCR (GSP3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>cpr-3</td>
<td>cpr-3/R3</td>
<td>cpr-3/R2</td>
<td>cpr-3/R1</td>
</tr>
<tr>
<td>cpr-4</td>
<td>cpr-4/R3</td>
<td>cpr-4/R2</td>
<td>cpr-4/R1</td>
</tr>
<tr>
<td>cpr-5</td>
<td>cpr-5/R3</td>
<td>cpr-5/R2</td>
<td>cpr-5/R1</td>
</tr>
<tr>
<td>cpr-6</td>
<td>cpr-6/R3</td>
<td>cpr-6/R2</td>
<td>cpr-6/R1</td>
</tr>
</tbody>
</table>

**Note:** The nucleotide sequences of these primers are listed in the appropriate table
above (Section 2.11.3.2).

Control Primers: GSP1: 5'-TTGTAATTCATTAAGCATTTCTGCC-3'
GSP2: 5'-CAUCAUCUACUGACATGGGAAGCCATCAGAC-3'

The Anchor Primer

**SpeI**

5'-CUACUACUACUGCCACGCGTCACTAGTACGGGIGGGIGGGGIGGGGIG-3'

UDG cloning site SalI anchor region

(Taken from the GibcoBRL 5' RACE protocol)
The Universal Primer

5'-AGGCCACGCGTGAATAGTACGGG-3'

The Universal Primer used was synthesised in the laboratory because of limited quantities of the GibcoBRL Universal Primer supplied with the kit. The Universal Primer used is essentially identical to the one supplied with the kit but lacks the sequences required for UDG cloning since this was not required for the cloning approach employed.

2.12.1. Reaction conditions of the nested PCR amplification steps

2.12.1.1. First round of nested PCR

The reaction conditions used to amplify the dC-tailed cDNA in the first round of nested PCR are given below:

<table>
<thead>
<tr>
<th>Reagent</th>
<th>Final Concentration</th>
</tr>
</thead>
<tbody>
<tr>
<td>Taq polymerase 10x buffer (Promega Corporation)</td>
<td>1x</td>
</tr>
<tr>
<td>Taq polymerase (Promega Corporation)</td>
<td>1U/10µl</td>
</tr>
<tr>
<td>Acetylated BSA (Promega Corporation)</td>
<td>100µg/ml</td>
</tr>
<tr>
<td>MgCl₂</td>
<td>2.5mM</td>
</tr>
<tr>
<td>dNTPs (Promega Corporation.)</td>
<td>200µM</td>
</tr>
<tr>
<td>Anchor Primer</td>
<td>40nM</td>
</tr>
<tr>
<td>pooled gene specific primers</td>
<td></td>
</tr>
<tr>
<td>for each of the four genes</td>
<td>100nM (each)</td>
</tr>
</tbody>
</table>

The reactions were made up in a final volume of 50µl using sdiH₂O. Each reaction was overlaid with an approximately equal volume of mineral oil and PCR performed using an Omnigene thermocycler (Hybaid Ltd.) using the following conditions:
<table>
<thead>
<tr>
<th>Step</th>
<th>Cycles</th>
<th>Temp(°C)</th>
<th>Time (min)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>94</td>
<td>3</td>
</tr>
<tr>
<td>2.1</td>
<td>35</td>
<td>57</td>
<td>0.5</td>
</tr>
<tr>
<td>2.2</td>
<td></td>
<td>72</td>
<td>2</td>
</tr>
<tr>
<td>2.3</td>
<td></td>
<td>94</td>
<td>0.5</td>
</tr>
<tr>
<td>3.1</td>
<td>1</td>
<td>57</td>
<td>2</td>
</tr>
<tr>
<td>3.2</td>
<td></td>
<td>72</td>
<td>10</td>
</tr>
</tbody>
</table>

2.12.1.2. Second round of nested PCR

The reaction conditions used to amplify the products from the first round of the nested PCR are given below. 1/1,000th of the amplified products from the first round PCR were amplified using the following conditions:

<table>
<thead>
<tr>
<th>Reagent</th>
<th>Final Concentration</th>
</tr>
</thead>
<tbody>
<tr>
<td>Taq polymerase 10x buffer (Promega Corporation)</td>
<td>1x</td>
</tr>
<tr>
<td>Taq polymerase (Promega Corporation)</td>
<td>1 unit/10μl</td>
</tr>
<tr>
<td>Acetylated BSA (Promega Corporation)</td>
<td>100μg/ml</td>
</tr>
<tr>
<td>MgCl₂</td>
<td>2.5mM</td>
</tr>
<tr>
<td>dNTPs (Promega Corporation)</td>
<td>200μM</td>
</tr>
<tr>
<td>Universal Primer</td>
<td>300nM</td>
</tr>
<tr>
<td>gene specific primer for</td>
<td>100nM</td>
</tr>
<tr>
<td>one of the four genes</td>
<td></td>
</tr>
</tbody>
</table>

The reactions were made up in a final volume of 50μl using sdiH₂O. Each reaction was overlaid with an approximately equal volume of mineral oil and PCR performed using an Omnigene thermocycler (Hybaid Ltd.) using the same conditions as for the first round of PCR.

2.13. Semi-quantitative reverse transcription polymerase chain reaction (s-q rtPCR)

Semi-quantitative reverse transcription Polymerase Chain Reaction (sq-rtPCR) was used to analyse the mRNA abundance of *cpr-3*, *cpr-4*, *cpr-5* and *cpr-6* through development. The 20 staged cDNA samples covering *C.elegans* larval and early adult
development were supplied by Dr. I.L. Johnstone. These samples were generated in the following manner. Embryos were isolated from *C. elegans* and allowed to hatch in the absence of an *E. coli* OP50 food source to produce larvae arrested in *L*₁, the resulting arrested *L*₁ larvae were seeded onto NGM agar plates with *E. coli* OP50 as a food source and grown at 25°C. Samples of approximately 2000 worms were taken at 2 hour intervals and total RNA extracted. This yielded 20 RNA samples covering the 4 larval stages and the first 10 hours of adult development. Samples from older adults were not taken because they contained embryos. A fraction (1/100<sup>th</sup>) of the RNA from each time point was converted to first strand cDNA using Superscript II reverse transcriptase (GibcoBRL) and random hexamers. Mixed stage embryo RNA was isolated as described in Section 2.7.2. First strand cDNA was prepared from 2μg of the mixed stage embryo total RNA using 4μg of random hexamers (Pharmacia) and Superscript II reverse transcriptase enzyme and buffer (GibcoBRL) according to the instructions supplied with the enzyme.

The cDNA from each time point was amplified simultaneously using two sets of PCR primers, one set specific to the test gene (*cpr-3*, *cpr-4*, *cpr-5* or *cpr-6*) and the second set specific to the internal control transcript, *ama-1*. The primers used to amplify each transcript were designed either to distinguish cDNA from contaminating genomic DNA, or to prevent amplification of genomic DNA. Three pools of cDNA from each developmental stage were amplified with 23 cycles of PCR. The number of PCR cycles was chosen empirically to ensure that reactants were still in excess once the cycles were complete. The products were electrophoresed through 1.2%-2.0% agarose gels, Southern blotted and probed simultaneously with probes specific to the target and *ama-1* genes. Probes were initially generated by 5′ end-labelling the same primers used for the PCR amplification step (see Section 2.8.7.2). In later experiments, the probes were obtained by PCR amplification of the *cpr-3*, *cpr-4*, *cpr-5* and *cpr-6* cDNA clones, or by PCR amplification from pooled *L*₁ cDNA in the case of *ama-1*, using the same primer pairs as for the s-q rtPCR and gel purification of PCR products of the correct size (see below).

After autoradiography to detect signal, regions of the membrane corresponding to the amplified cDNA products of each gene were excised and radioactivity measured by counting in scintillant using a Beckman LS801 scintillation counter. The relative
abundance of the test gene was then expressed as a ratio of the signal from the test gene to that of *ama-1* for each developmental stage. The mean and standard deviation of the ratios from each set of triplicate results were calculated.

2.13.1. PCR primers used for s-q rtPCR

The PCR primers used for the s-q rtPCR experiments for each of the four genes are described below:

<table>
<thead>
<tr>
<th>Primer Pair</th>
<th>Primer Name</th>
<th>Primer Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>cpr-3</em> primer pair</td>
<td><em>cpr-3/1</em></td>
<td>listed above</td>
</tr>
<tr>
<td></td>
<td><em>cpr-3/R3</em></td>
<td>listed above</td>
</tr>
<tr>
<td><em>cpr-4</em> primer pair</td>
<td><em>cpr-4/0.5</em></td>
<td>listed above</td>
</tr>
<tr>
<td></td>
<td><em>cpr-4/R3</em></td>
<td>listed above</td>
</tr>
<tr>
<td><em>cpr-5</em> primer pair</td>
<td><em>cpr-5/1</em></td>
<td>listed above</td>
</tr>
<tr>
<td></td>
<td><em>cpr-5/R3</em></td>
<td>listed above</td>
</tr>
<tr>
<td><em>cpr-6</em> primer pair</td>
<td><em>cpr-6/2</em></td>
<td>listed above</td>
</tr>
<tr>
<td></td>
<td><em>cpr-6/R4</em></td>
<td>listed above</td>
</tr>
<tr>
<td><em>ama-1</em> Primer Pair</td>
<td><em>ama-1/1</em></td>
<td>5'-TTCCAAGCGCCGCTGCGCATTTGTCTC-3'</td>
</tr>
<tr>
<td></td>
<td><em>ama-1/2</em></td>
<td>5'-CAGAATTTCCAGCACTCGAGGAGCGGA-3'</td>
</tr>
</tbody>
</table>

The PCR primers used for the s-q rtPCR analysis of *cpr-4* proved to be non-specific and therefore a new PCR primer pair was designed. The names and nucleotide sequences of these primers are given below, in addition to the position of the 5' end and 3' end bases of these primers with respect to the ATG initiation codon of *cpr-4* (Chapter 4, Figure 5.12).

<table>
<thead>
<tr>
<th>Primer Name</th>
<th>Primer Sequence</th>
<th>Primer Position</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>5' Base</td>
</tr>
<tr>
<td><em>cpr-4/0.5PCR</em></td>
<td>5'-GCTCTATCTTTGCTATTTGCTC TTTTACAAAAATGAAATAC-3'</td>
<td>-31</td>
</tr>
<tr>
<td><em>cpr-4/R1.5PCR</em></td>
<td>5'-GGCTCCGTGGAGGCGATGCAG-3'</td>
<td>+436</td>
</tr>
</tbody>
</table>
2.13.2. The PCR reaction conditions used for s-q rtPCR

The same reaction conditions were used in the PCR amplification step of the s-q rtPCR experiments for each of the four genes. A general mix was prepared which contained the following:

<table>
<thead>
<tr>
<th>Reagent</th>
<th>Final Concentration</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Taq</em> polymerase 10x buffer (Promega Corporation)</td>
<td>1x</td>
</tr>
<tr>
<td>Acetylated BSA (Promega Corporation)</td>
<td>100µg/ml</td>
</tr>
<tr>
<td>MgCl2</td>
<td>1.5mM</td>
</tr>
<tr>
<td>dNTPs (Promega Corporation)</td>
<td>200µM</td>
</tr>
<tr>
<td><em>ama-1</em> primer pair</td>
<td>3ng/µl</td>
</tr>
<tr>
<td>gene specific primer pair</td>
<td>3ng/µl</td>
</tr>
</tbody>
</table>

For each gene, sufficient volumes of the above mix was prepared for all the required PCR reactions. The mix was heated to 70°C for 5 min and then *Taq* DNA polymerase (Promega Corporation) was added to a final concentration of 1U/25 µl. The reaction mixture was mixed and 24 µl dispensed to a number of tubes each containing 1 µl of the cDNA sample generated from each RNA time point, prewarmed to 80°C. Aerosol resistant tips were used for this to avoid cross contamination between samples. After the reaction mix had been added to each tube, and mixed with the template cDNA, an approximately equal volume of mineral oil was added to each tube to prevent evaporation and the PCR cycle profile activated immediately. In later experiments, the reactions were performed in a final volume of 50 µl. The reaction conditions were identical and 48 µl of reaction mix was added to 2 µl of each cDNA sample, for each time point, giving the same ratio of template to reaction mix as for the 25 µl reactions. In all cases the PCR amplification step of the s-q rtPCR experiments were performed using an Omnigene thermocycler (Hybaid Ltd.). The PCR cycle profile used for the s-q rtPCR experiments for *cpr-3*, *cpr-5* and *cpr-6* is listed below:
<table>
<thead>
<tr>
<th>Step</th>
<th>Cycles</th>
<th>Temp(°C)</th>
<th>Time (min)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>94</td>
<td>3</td>
</tr>
<tr>
<td>2.1</td>
<td>22</td>
<td>60</td>
<td>0.5</td>
</tr>
<tr>
<td>2.2</td>
<td></td>
<td>72</td>
<td>2</td>
</tr>
<tr>
<td>2.3</td>
<td></td>
<td>94</td>
<td>0.5</td>
</tr>
<tr>
<td>3.1</td>
<td>1</td>
<td>60</td>
<td>2</td>
</tr>
<tr>
<td>3.2</td>
<td></td>
<td>72</td>
<td>10</td>
</tr>
</tbody>
</table>

Note: For cpr-4, the temperature of the annealing steps, 2.1 and 3.1, were reduced from 60°C to 55°C.

2.13.3. PCR amplification of sequences required to probe the s-q rtPCR Southern blots

The same reaction conditions, primers and cycle profiles used for the PCR amplification step of s-q rtPCR were used to amplify the sequences required for probing the s-q rtPCR Southern Blots with the following alterations:

i) To obtain probes for cpr-3, cpr-4, cpr-5 and cpr-6, approximately 200ng of the cloned cDNAs for these genes (in pBluescript II KS+) were used as template. For ama-1, 1μl of an unknown concentration of cDNA pooled from the 6 L1 stage time points obtained from Dr I.L.Johnstone was used as template.

ii) The amount of Taq DNA polymerase used was increased to 1unit/10μl

iii) The number of cycles of steps 2.1.2.2 and 2.3 were increased from 22 to 35 cycles.

2.14. Colony PCR

Colony PCR was used to screen cloning steps performed during the generation of the cpr-3::lacZ, cpr-4::lacZ, cpr-5::lacZ and cpr-6::lacZ fusion constructs. The same reaction conditions and PCR cycle profiles were used in all cases and are listed below:
Reagents | Final Concentration
--- | ---
*Taq* polymerase 10x buffer (Promega Corporation) | 1x
*Taq* DNA polymerase (Promega Corporation) | 1 unit/25μl
Acetylated BSA (Promega Corporation) | 100μg/ml
MgCl₂ | 1.5mM
dNTPs (Promega Corporation) | 200μM
Primer A (listed below) | 4ng/μl
Primer B (listed below) | 4ng/μl

**Table i. Primer A**

<table>
<thead>
<tr>
<th>Primer Name</th>
<th>Primer Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>M13 Reverse PCR Primer</td>
<td>listed above (Section 2.11.3.3)</td>
</tr>
<tr>
<td>cpr-3 Primer ‘X’ (cpr-3/5'FL)</td>
<td>listed above (Section 2.11.3.2, Table i)</td>
</tr>
<tr>
<td>cpr-4 Primer ‘X’(cpr-4/5'FL)</td>
<td>listed above (Section 2.11.3.2, Table ii)</td>
</tr>
<tr>
<td>cpr-5 Primer ‘X’(cpr-5/5'FL)</td>
<td>listed above (Section 2.11.3.2, Table iii)</td>
</tr>
<tr>
<td>cpr-6 Primer ‘X’(cpr-6/5'FL)</td>
<td>listed above (Section 2.11.3.2, Table iv)</td>
</tr>
</tbody>
</table>

**Table ii. Primer B**

<table>
<thead>
<tr>
<th>Primer Name</th>
<th>Primer Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>M13 Forward PCR Primer</td>
<td>5’-GTTGTAAAAACGACGCCAGT-3’</td>
</tr>
<tr>
<td>lacZ/2</td>
<td>5’-TCGCCATTCAGGCTGCGCAACTGTT-3’</td>
</tr>
<tr>
<td>SYNINT</td>
<td>listed above (Section 2.11.3.3)</td>
</tr>
<tr>
<td>NUCLOC</td>
<td>listed above (Section 2.11.3.3)</td>
</tr>
</tbody>
</table>

25μl of the reaction mixture was dispensed to as many 0.5ml eppendorf tubes as were required for the colony PCR screen. Isolated colonies from transformation plates or from streaked stock plates were picked using a wooden tooth pick. The pick was dipped in one of the eppendorf tubes, mixed and then dropped into 5ml of L-broth supplemented with the appropriate antibiotic to generate an overnight culture. The PCR reactions were overlaid with an approximately equal volume of mineral oil and PCR amplified using the following PCR cycle profiles:
<table>
<thead>
<tr>
<th>Step</th>
<th>Cycles</th>
<th>Temp(°C)</th>
<th>Time (min)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>94</td>
<td>3</td>
</tr>
<tr>
<td>2.1</td>
<td>20</td>
<td>55</td>
<td>0.5</td>
</tr>
<tr>
<td>2.2</td>
<td></td>
<td>72</td>
<td>3</td>
</tr>
<tr>
<td>2.3</td>
<td></td>
<td>94</td>
<td>0.5</td>
</tr>
<tr>
<td>3.1</td>
<td>1</td>
<td>55</td>
<td>3</td>
</tr>
<tr>
<td>3.2</td>
<td></td>
<td>72</td>
<td>15</td>
</tr>
</tbody>
</table>

Note: For colony PCR involving the large 5’ flank insert of cpr-6, the extension time of step 2.2 was increased to 4.5min. For colony PCR using the cpr-5 Primer ‘X’, the annealing temperatures of steps 2.1 and 3.1 were reduced to 50°C.

2.15. High fidelity PCR

The 5’ flank linkers required for generation of the lacZ fusion constructs cpr-3::lacZ, cpr-4::lacZ, cpr-5::lacZ and cpr-6::lacZ, were obtained using high fidelity PCR to reduce the probability of introducing PCR generated mutations. The principle differences to other PCR reactions were the use of a thermostable DNA polymerase with 3’-5’ exonuclease activity (proof-reading activity) and reduction of the total number of PCR cycles. The conditions used were as follows

**Reagents**

*Pwo DNA Polymerase*
(Boehringer Mannheim UK Ltd., Lewes, E.Sussex, UK)

*Pwo DNA Polymerase 10x Buffer*
Including 20mM MgSO₄
(Boehringer Mannheim UK Ltd.)

Dialysed BSA (Pharmacia)

dNTPs (Promega Corporation)

Primer ‘X’ (listed below)

Primer ‘Y’ (listed below)

Template DNA (listed below)

**Final Concentration**

1unit/40µl

1x

100µg/ml

200µM

3ng/µl

3ng/µl

1µg
Each reaction was made up in a final volume of 100μl using sdiH2O and overlaid with approximately 100μl of mineral oil. The PCR reactions were then amplified using the following PCR cycle profiles:

<table>
<thead>
<tr>
<th>Step</th>
<th>Cycles</th>
<th>Temp (°C)</th>
<th>Time (min)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>94</td>
<td>3</td>
</tr>
<tr>
<td>2.1</td>
<td>14</td>
<td>50</td>
<td>0.5</td>
</tr>
<tr>
<td>2.2</td>
<td></td>
<td>72</td>
<td>1</td>
</tr>
<tr>
<td>2.3</td>
<td></td>
<td>94</td>
<td>0.5</td>
</tr>
<tr>
<td>3.1</td>
<td>1</td>
<td>50</td>
<td>2</td>
</tr>
<tr>
<td>3.2</td>
<td></td>
<td>72</td>
<td>10</td>
</tr>
</tbody>
</table>

### Nucleotide Sequences of Primer ‘X’ and ‘Y’ for Each Gene

<table>
<thead>
<tr>
<th>Gene</th>
<th>Cloned Template</th>
<th>Primer ‘X’</th>
<th>Primer ‘Y’</th>
</tr>
</thead>
<tbody>
<tr>
<td>cpr-3</td>
<td>CL#034/1.7/KS+</td>
<td>cpr-3/5’ FL</td>
<td>5’-TAGAGGGATCCGCCTAAAT GTAGAAAGATATGG-3’</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>cpr-4</td>
<td>ZK1055/2.8/KS+</td>
<td>cpr-4/5’ FL</td>
<td>5’-AGAACCAGGATTTCTTC AGACAATTACATA-3’</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>cpr-5</td>
<td>W02B2/3/KS+</td>
<td>cpr-5/5’ FL</td>
<td>5’-GAGACCCGGACATTATG AGAAGTGTCGCG-3’</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>cpr-6</td>
<td>C25B8/3.8/KS+</td>
<td>cpr-6/5’ FL</td>
<td>5’-GAGCAGGATCTTCTAGAA TTCAATTTTCTCAA-3’</td>
</tr>
</tbody>
</table>

Notes: The nucleotide sequences of Primer ‘X’ for each gene are given in the appropriate table of Section 2.11.3.2. The nucleotide sequences representing restriction endonuclease sites (BamHI for cpr-3 and cpr-6, XmaI for cpr-4 and cpr-5) engineered into the 5’ end of Primer ‘Y’ for each gene are highlighted in bold.

2.16. **PCR amplification of the 5’ ends of the cloned cDNAs of cpr-3, cpr-4, cpr-5 and cpr-6**

The 5’ ends of the four cDNA clones were required for probing Southern blots during restriction endonuclease mapping of the genomic clones for each of the four genes, in order to determine the orientation of each gene within each clone. The 5’ ends of the cDNA clones were generated using PCR and the following conditions:
Reagents

*Taq* polymerase 10x buffer (Promega Corporation) 1x
*Taq* DNA polymerase (Promega Corporation) 1 unit/10μl
Acetylated BSA (Promega Corporation) 100μg/ml
MgCl₂ 2.5mM
dNTPs (Promega Corporation) 200μM
M13 Forward PCR Primer 2ng/μl
Gene Specific Primer (listed below) 2ng/μl
Template cDNA clone (listed below) 200ng

**Template**

cpr-3 cDNA clone (cm12b6) cpr-3/R1 (Section 2.11.3.2, Table i)
cpr-4 cDNA clone (cm14e3) cpr-4/R1 (Section 2.11.3.2, Table ii)
cpr-5 cDNA clone (cm04d10) cpr-5/R1 (Section 2.11.3.2, Table iii)
cpr-5 cDNA clone (cm01a5) cpr-6/R1 (Section 2.11.3.2, Table iv)

The reactions were made up in a final volume of 50μl and overlaid with an approximately equal volume of mineral oil prior to thermal cycling. The PCR cycle profiles used were as follows:

<table>
<thead>
<tr>
<th>Step</th>
<th>Cycles</th>
<th>Temp(°C)</th>
<th>Time (min)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>94</td>
<td>0.5</td>
</tr>
<tr>
<td>2.1</td>
<td>35</td>
<td>60</td>
<td>0.5</td>
</tr>
<tr>
<td>2.2</td>
<td></td>
<td>72</td>
<td>1</td>
</tr>
<tr>
<td>2.3</td>
<td></td>
<td>94</td>
<td>0.5</td>
</tr>
<tr>
<td>3.1</td>
<td>1</td>
<td>60</td>
<td>2</td>
</tr>
<tr>
<td>3.2</td>
<td></td>
<td>72</td>
<td>10</td>
</tr>
</tbody>
</table>

2.17. Generation of transgenic *C.elegans* lines and staining for β-Galactosidase activity

2.17.1. Preparation of materials and worms for microinjection

Plasmid DNA for microinjection was extracted and purified using the Qiagen Plasmid Maxi or Plasmid Midi DNA kits (Qiagen Inc.) and was resuspended in TE. Each of the four *lacZ* fusion constructs, *cpr-3::lacZ*, *cpr-4::lacZ*, *cpr-5::lacZ* and *cpr-6::lacZ* fusion were diluted to 400μg/ml with sterile distilled water. The pRF4 marker
plasmid, carrying the dominant *rol-6* (*su1006*) mutation was diluted to 200\(\mu\)g/ml in TE. Each of the diluted *lacZ* fusion constructs were mixed 1:1 with the diluted pRF4 plasmid to generate a final injection solution of 200\(\mu\)g/ml of *lacZ* fusion construct and 100\(\mu\)g/ml pRF4 marker plasmid.

Needles for microinjection were prepared from Aluminum Silicate glass capillaries (Clark Electromedical Instruments, Pangbourne, Reading, UK) which possessed glass filaments along their internal length to aid loading of the injection solution to the tip of the needle. Needles were pulled to the desired shape using a model 773 electrode puller (Campden Instruments Ltd., Sileby, Loughborough, UK). Immediately before use, the injection solution was centrifuged at 15,000rpm in a microcentrifuge (Heraeus Equipment Ltd.) and loaded into hand pulled borosilicate capillary tubes (Clark Electromedical Instruments) by drawing the injection solution up using a mouth pipettor. The DNA was then loaded into the needles by inserting the hand pulled capillary tube down the wide bore of the needle and depositing the injection solution just behind the needle tip, using gentle pressure with a mouth pipettor (see Section 2.17.2)

Worms to be injected were immobilised on dried agarose pads under paraffin oil essentially as described by A.Fire (Fire, 1986). The dried agarose pads were prepared on 60mm x 22mm glass coverslips (#1), by placing a single drop of a 3% agarose solution onto the coverslip, placing a 22mm x 22mm glass coverslip (#1) on the drop and applying gentle pressure. The smaller coverslip was removed and the agarose plug dried by baking at 80°C for 10min.

Young adult N2 Bristol *C.elegans* hermaphrodite worms were transferred to a partially desiccated NGM agar plate prior to injection. When the injection apparatus was ready, a sharpened platinum wire pick was used to transfer 5-15 of these worms to a dried agarose plug, overlaid with 2-3 drops of paraffin. The picked worms were allowed to touch the dried agarose plug and the pick was gently removed once a portion of the worm’s body had adhered to the surface of the dried agarose. The immobilised worms were placed on the microscope stage and the injection procedure started immediately.
2.17.2. Microinjection procedure

The microinjection procedure and apparatus are essentially the same as that described by C. Mello et al. (1991). Needles containing the injection solution were loaded into the microtool collar of a Narishige MO-202 Joystick Hydraulic Micromanipulator (Narishige Co Ltd., Tokyo 157, Japan) and were pressurised using an attached bottled nitrogen source. A pressure regulator and solenoid valve attached to the nitrogen cylinder were used to obtain the desired rate of flow of the injection solution. Pressure was activated by means of a foot operated switch linked to the solenoid valve.

Each large coverslip, carrying worms immobilised on the agarose plugs, was taped to the free sliding oil cushioned stage of a Zeiss Axiovert 100 inverted microscope fitted with Nomarski optics. Microinjections were performed at 400x magnification using a Zeiss Achromplan 40x air objective lens. Pressure was applied to the needle containing the injection solution using the foot operated switch. The micromanipulator was used to open the sealed needle by repeatedly touching the tip of the needle onto a piece of capillary tube placed on the agarose plug until a suitably high flow rate was observed, indicating that the needle possessed a sufficiently large bore to reduce blocking. Injections were performed by first focusing on the grainy cytoplasm in the centre of the distal gonad. The needle was then brought into the same focal plane and the microscope stage moved by hand to push the worm into the needle tip. Once the needle was touching the surface of the worm, the microscope stage was tapped and this usually resulted in the needle penetrating the cuticle and gonad. Nitrogen pressure was then applied to force the injection solution into the gonad immediately after penetration. The needle was rapidly withdrawn from the worm without reducing the nitrogen pressure, once the gonad had swollen visibly.

After injection, the worms were recovered by adding a drop of 1xM9 solution onto the agarose pad, causing the worms to float off the pad. The injected worms were then collected from the drop using a platinum wire pick and approximately 10 worms seeded onto a fresh NGM agar plate supporting an E.coli OP50 food source, by floating the worms onto a drop of sterile 1xM9 solution. The injected worms were incubated for 3-4 days at 20°C until F1 progeny with the rol-6 phenotype were observed.
2.17.3. Isolation of transgenic lines

F1 progeny from the injected worms which exhibited the *rol-6* phenotype, conferred by the pRF4 marker plasmid, were seeded to fresh NGM agar plates supporting an *E.coli* OP50 food source (10-20 worms/plate). These plates were incubated for 3-4 days at 20°C until F2 progeny with the *rol-6* phenotype were observed. At least 3 F2 roller progeny were picked from each F1 plate that transmitted the *rol-6* phenotype to the F2 progeny and these worms were seeded onto fresh NGM agar plates supporting an *E.coli* OP50 food source (1 worm/plate). This process generated at least 3 populations of worms derived from a single animal from each original plate of transmitting F1 progeny. Of the populations generated from each F1 plate, the one with the greatest fraction of worms with the *rol-6* phenotype was selected as a source for the transmitting line. In this manner, at least 2 independent transmitting lines were generated for each injected *lacZ* fusion construct. These lines transmitted the roller phenotype to 10% - 90% of their offspring.

2.17.4. Fixing of transgenic worms

The worms were fixed essentially as described by J.M.Young and I.A.Hope (Young and Hope, 1993). 10-20 worms with the roller phenotype were seeded onto large (9cm) NGM plates seeded with *E.coli* OP50 and incubated until the bacterial lawn was almost cleared (usually 4-5 days). The worms were washed off the plates using 1xM9 and pelleted at 5,000rpm for 10 seconds in a Heraeus microfuge. Approximately 100 worms in 10μl were placed on a subbed microscope slide (prepared by dipping the slide into a 0.01% poly-L-lysine solution and baking at 80°C for 10min) and overlaid with a 60mm x 22mm coverslip (#1). The slide was snap frozen on an aluminium block standing in dry ice and the cover slip removed with a scalpel blade. The slide was immediately immersed in methanol at -20°C and incubated at-20°C for 5min. The slide was transferred to acetone at -20°C and incubation continued at -20°C for a further 5min. The slides were air dried (worm side up) for 1 hour at room temperature using a large glass plate to act as a heat sink.
2.17.5. Staining of transgenic worms

Staining for β-galactosidase activity was performed using the substrate X-gal essentially as described by A.Fire et al (1990). 40μl of sensitive staining solution (see below), prewarmed to 42°C, was added to the slide carrying the fixed worms, also prewarmed to 42°C, and overlaid with a 60mm x 22mm coverslip (#1). The coverslip was sealed to the slide using nail varnish and incubated at 42°C in a humidified environment until staining nuclei were visible (5min - 24 hours). The worms were examined at 100x, 200x and 400x magnification using a Zeiss Axioplan microscope and Zeiss Plan-Neofluar 10x/0.3, 20x/0.5 and 40x/0.75 air objectives, respectively. Photographs were taken using a Zeiss MC100 Spot camera and Kodak Ektachrome 64T colour reversal or Kodak Ektapress Gold II 100 colour negative 35mm film (Kodak Ltd.).

Sensitive Staining Solution

0.2M sodium phosphate buffer pH7.5
1mM MgCl₂
10mM redox buffer (50mM potassium ferricyanide, 50mM potassium ferrocyanide)
0.004% SDS
0.3% X-Gal (from a 30% (w/v) stock in dimethyl formamide, prepared fresh)

2.18. Computer programs

All DNA sequence computational data analyses were performed using programs from the GCG package (according to the Program Manual for the Wisconsin Package, version 8, (1994), Genetics Computer Group, 575 Science Drive, Madison, Wisconsin, USA, 53711). All phylogenetic analyses were performed using programs from the PHYLIP package version 3.5c (Felsenstein, 1993) according to the program manuals supplied with the package. The ACEDB C.elegans database for Macintosh computers (MACACE) was used for all C.elegans genetic and physical map computer analyses.
Chapter 3
Chapter 3

Cloning Four Cathepsin B-like Genes from C. elegans

3.1. Introduction

3.1.1. The C. elegans physical map

In order to sequence a genome, it is necessary to generate an ordered physical map of clones covering that genome so that sequence data obtained from these clones can be accurately linked together. Furthermore, good alignment between such physical maps and their corresponding genetic maps are required in order to allow genes identified by sequencing to be accurately aligned to the genetic map.

Currently, the physical map of C. elegans comprises overlapping cosmid, lambda and yeast artificial chromosome (YAC) clones covering over 95 Mb (95%) of the genome. Initially, the physical map was assembled from cosmid and lambda clones using a restriction enzyme fingerprinting method to link overlapping clones into large sets known as “contigs” (Coulson et al., 1986). This approach resulted in 90-95% of the C. elegans genome being cloned into 17,500 cosmids assembled into 700 contigs. The contigs could not be joined to each other because the linking clones required to do this were either not represented or severely under represented in the cosmid libraries. In order to overcome this problem, a YAC library of C. elegans genomic DNA was generated and a hybridisation strategy used to link the YAC and cosmid clones (Coulson et al., 1988). Two sets of ordered grids were prepared; one of random YAC clones, representing approximately 2.5 genome equivalents, and the other of cosmids, which represented the contigs and unattached clones. Cosmids from the end of contigs were used to probe the YAC grid. YAC clones selected using this approach were isolated from host chromosomes by pulsed field gel electrophoresis and used to probe the cosmid grid. This proved to be a very effective approach for linking contigs together, with the number of contigs being reduced from 700 to 346 within seven months of the initiation of this project. This approach also facilitated the generation of a lower resolution physical map consisting of 958 ordered, overlapping YAC clones known as the YAC
grid (Coulson et al., 1991), since the hybridisation technique used also revealed the precise position and extent of the YAC clones with respect to the cosmid clones. At present, the majority of the physical map of C. elegans consists of a high resolution map of ordered cosmid and lambda clones covered by a lower resolution map of YAC clones represented on the YAC grid.

The physical and genetic maps of C. elegans are also well aligned, with over 90 Mb being positioned along the chromosomes using genetically mapped sequences and in situ hybridisation techniques (Sulston et al., 1992). The accuracy of the alignment continues to increase as more genetically mapped sequences are determined.

3.1.2. The C. elegans genome sequencing project

The genome sequencing project started near the centre of chromosome III (Sulston et al., 1992) because there was good cosmid coverage over several megabases of sequence and because the genetic map suggested that the central region of C. elegans chromosomes are gene rich. In the last report, sequences from 732 clones have been obtained, resulting in 22.5 Mb of finished sequence and an additional 15 Mb of sequence at various stages of completion (The C. elegans Genome Sequencing Consortium, pers. comm.). These sequences are primarily from the central regions of chromosomes II, III, IV and X. Once finished sequences have been obtained, the program GENEFINDER is used to find possible genes. These potential genes are then compared to sequences already in the public domain databases, such as GenBank, using the BLASTX program. Such analysis has revealed that approximately 46% of these predicted genes show significant similarities to sequences already in the databases. These predicted genes occur at an average density of one per 4.8 kb, for chromosomes II and III, or 6 kb, for the X chromosome (The C. elegans Genome Sequencing Consortium, pers. comm.). Based on these data, there are currently estimated to be approximately 13,526 (+/- 500) genes in C. elegans (The C. elegans Genome Sequencing Consortium, pers. comm.).

As a complementary approach to the C. elegans genome sequencing project, several groups are sequencing and mapping randomly selected cDNA clones from C. elegans. These projects generate partial sequence from cDNA clones which have been mapped to the YAC grid to produce sequence tags. These sequence tags can then be
used to predict coding regions from the genomic sequence generated by the genome sequencing project. The initial results from two laboratories yielded approximately 1600 different cDNA sequences. McCombie et al (1992) isolated 585 random clones from a mixed stage cDNA library. These clones were partially sequenced from the 5' and/or 3' ends to produce 720 expressed sequence tags (ESTs). These ESTs were subsequently compared to sequences in GenBank. After redundancy was taken into consideration (where one gene is represented by more than one EST), 437 nuclear genes were identified. R.Waterston et al (1992) used a sorted cDNA library as a source of cDNA clones for random sequencing. The cDNA clones from this library were isolated in several cycles in which clones that hybridised to previously selected clones were discarded. This process generates a library that is partially normalised for expression levels, since isolation of multiple cDNA clones from the same gene will be selected against. For the same reason, this pre-selection screens out most highly homologous members of the same gene family. This screen has yielded sequences from 1,517 clones which appear to represent 1,194 individual genes. Comparison of the different cDNA sequences to sequences in the Genbank and EMBL databases revealed very similar results from each of these two screens. Each group identified 15 previously described *C.elegans* genes. Of the remaining sequences, both groups found that 25-30% of the cDNA sequences showed homology to previously isolated genes in the databases. In both screens, the majority of the clones obtained represent entirely novel sequences which will be very useful for determining coding sequence from the genome sequencing project. Both groups are currently mapping the partially sequenced cDNA clones to the *C.elegans* physical map (McCombie et al., 1992; Waterston et al., 1992). More recently, Yuji Kohara has implemented a project aimed at isolating all the cDNA species of *C.elegans* (Yuji Kohara, pers. comm.). In this project, the cDNA clones are not only being partially sequenced and physically mapped using the YAC grid, but also used to analyse expression patterns in whole mount embryos using a high throughput multi-well *in situ* hybridisation approach. Clearly, such projects will generate information that is not only useful for identifying genes from sequence generated by the genome sequencing project, but is also useful for helping to understand the biological roles of those genes identified.
3.1.3. Additional uses of the *C. elegans* physical map

The virtually complete physical map of *C. elegans* is not only an essential prerequisite for sequencing of the genome but is also an extremely useful tool in its own right. The physical map position of most cloned sequences can be determined rapidly by hybridisation to the YAC grid. This information may then be used to facilitate the rapid isolation of genomic clones corresponding to the sequences used to probe the YAC grid since cosmid and lambda clones can be selected which subtend the region covered by the hybridising YAC clones of the YAC grid, so avoiding the need to screen an entire cosmid or lambda library. The good alignment of the physical and genetic maps may also yield candidate mutant phenotypes for the genomic clone isolated. Furthermore, the ability to transform *C. elegans* allows the use of transgenic technology to determine whether an isolated genomic clone is capable of repairing such candidate mutant phenotypes. Thus the physical map of *C. elegans*, and its alignment to the genetic map, provides a potential route for rapidly cloning and mapping genes, some of which may be defined by mutations. This feature is especially useful since all the partially sequenced and mapped cDNA clones, generated in the screens discussed in Section 3.1.2, are being made available to researchers.

3.2. Results

3.2.1. Obtaining four cDNA clones encoding proteins with homology to cathepsin B

During their survey of expressed sequences from *C. elegans*, R. Waterston et al (1992) isolated and partially sequenced eight cDNA clones which showed significant homology to cathepsin B-like sequences in the public domain databases. A comparison of the partial sequence using the GAP program of the GCG package revealed that these cDNAs fell into four classes and were probably derived from four different genes. Between classes, the cDNA clones showed 37.8 - 62.6% identity. Within each class, the degree of identity rose dramatically to 91.6-100%. In the latter case, the lack of complete identity between some clones of the same class could be attributed to
sequencing errors since all the clones were only sequenced on one strand. Indeed, the
type of errors observed were consistent with this assumption, being primarily base
insertions or deletions, resulting in frameshifts and disruption of the open reading frame
for each of the clones, or a result of nucleotide residues whose base composition could
not be determined, resulting in residues denoted as ‘N’. Accordingly, four clones named
cm12b6, cm14e3, cm04d10 and cm01a5, representing each of the four classes, were
obtained from J.Sulston and A.Coulson for further analysis. Before any of the four
clones were analysed further, the pRATII plasmids containing the cDNA inserts of these
four clones were excised from the lambda SHLX2 vector and transfected into E.coli
XL1-blue (see Chapter 2, Section 2.10). E.coli XL1-blue cells containing the plasmids
were selected using the ampicillin resistance marker. The cDNA inserts from cm12b6,
cm14e3, cm04d10 and cm01a5 in pRATII were subsequently subcloned into pBluescript
II KS+ as ApaI/SacI fragments for further characterisation.

The ApaI/SacI cDNA inserts of the four selected cDNA clones, cm12b6, cm14e3, cm04d10 and cm01a5, were all used as probes to isolate genomic clones. Subsequent sequence analysis of the genomic clones (Chapter 4) revealed each to represent a distinct cathepsin B-like gene. These genes were named cpr-3, cpr-4, cpr-5 and cpr-6. A comparison of these genes to the cDNA sequence of cm12b6, cm14e3, cm04d10 and cm01a5 revealed complete base identity, interrupted by introns. Therefore the genes corresponding to the cDNA clones, cm12b6, cm14e3, cm04d10 and cm01a5, are subsequently referred to as cpr-3, cpr-4, cpr-5 and cpr-6 respectively.

3.2.2. Confirming the physical map positions of cpr-3, cpr-4, cpr-5 and cpr-6

The results obtained from the physical mapping of a representative clone from
each of the four classes of cathepsin B-like cDNA clones suggested that the gene, or
genes, corresponding to each class of clone were located at distinct loci, with each
cDNA clone hybridising strongly to a distinct region of the YAC grid (R.Waterston pers.
comm.). However, several cDNA clones also hybridised weakly to additional regions of
the grid. To confirm the preliminary YAC grid location of the four genes provided by
R.Waterston, yeast strains carrying the YAC clones to which the four cDNA clones
hybridised on the YAC grid were obtained from J.Sulston and A.Coulson for Southern
blot analysis. Total genomic DNA was extracted from yeast strains carrying the YACs; Y69A3, Y7E11, Y48G5, Y50D4, Y55B10 and Y43B9. The restriction endonucleases BamHI and EcoRI were used to digest 1µg of yeast genomic DNA, along with 5µg of wild type C.elegans Bristol N2 genomic DNA as a control. The digested DNA from each yeast strain was electrophoresed through a 0.8% agarose gel, with digested C.elegans genomic DNA in adjacent lanes. The DNA was Southern blotted to Hybond N (Amersham International plc), hybridised at medium stringency to the appropriate gel purified cDNA probe (radiolabeled by random priming) and washed twice at high stringency (Chapter 2, Section 2.8.8.3) prior to autoradiography.

According to the preliminary physical map data from R.Waterston; the genes cpr-3 and cpr-4, corresponding to the cDNA clones cm12b6 and cm14e3, are contained within the YAC clones Y69A3 and Y7E11, respectively. The gene cpr-5, corresponding to the cDNA clone cm04d10, is contained within the YAC clones Y50D4 but also shows weak hybridisation to a second locus covered by Y48G5. The gene cpr-6, corresponding to the cDNA clone cm01a5, is contained within two overlapping YAC clones, Y43B9 and Y55B10. The results of each Southern blot analysis (Figures 3.1, 3.2, 3.3 and 3.4) indicate that these preliminary map positions are correct. Each cDNA clone produces strong signals and very similar hybridisation patterns when used to probe the control C.elegans genomic DNA and the yeast genomic DNA containing the appropriate YAC clone. Thus, the cDNA clones for cpr-3, cpr-4 and cpr-5 (cm12b6, cm14e3 and cm04d10) hybridise to Y69A3 (Figure 3.1), Y7E11 (Figure 3.2) and Y50D4 (Figure 3.3A), respectively, while the cDNA clone for cpr-6 (cm01a5) hybridises to the two overlapping clones, Y55B10 and Y43B9 (Figure 3.4A and 3.4B). The cDNA clone for cpr-5 (cm04d10) only produces weak signals when hybridised to Y48G5 (Figure 3.3B). The same pattern of weakly hybridising fragments is observed with EcoRI digested yeast DNA (lane 4) and control C.elegans DNA (lane 2), suggesting that a gene with some homology to the cm04d10 cDNA clone of cpr-5 may be contained within Y48G5. However, no common pattern of hybridising fragments is observed for the BamHI digested yeast genomic DNA (lane 3) and control C.elegans genomic DNA (lane 1) suggesting that this weak hybridisation is not reproducible with other digests. Together these data indicate that the cpr-5 gene is not present at a second locus outwith the region covered by the Y50D4 YAC clone, but that there may be a gene in the region
covered by the Y48G5 YAC clone which shows slight homology to the cm04d10 cDNA clone of *cpr-5*.

### 3.2.3. Analysis of the Southern blots of *C. elegans* genomic DNA and yeast DNA

The cDNA clones produce very similar, if not identical, hybridisation patterns when used to probe the digested control *C. elegans* genomic DNA (lanes 1 and 2, Figures 3.1, 3.2, 3.3 and 3.4) and yeast genomic DNA containing the appropriate YAC clones (lanes 3 and 4, Figures 3.1, 3.2, 3.3 and 3.4). Furthermore, the patterns of hybridisation observed for each of the four genes are in agreement with the patterns predicted from *BamHI* and *EcoRI* restriction endonuclease sites revealed after sequencing the genomic clones of *cpr-3, cpr-4, cpr-5* and *cpr-6*. In addition to confirming the preliminary physical map data from R.Waterston, these data suggest that the genomic sequences corresponding to each of the cDNA clones are contained completely within the appropriate YAC clones. Furthermore, these data demonstrate the integrity of the genomic sequences corresponding to the four cDNA clones within each of the appropriate YAC clones. The hybridisation patterns observed for each gene will be discussed in more detail in the following sections.

#### 3.2.3.1. The hybridisation pattern for *cpr-3*

The nucleotide sequence of *cpr-3* revealed the presence of a single *BamHI* site towards the 5’ end of the coding region of this gene, but no *EcoRI* site. Thus, *BamHI* digests of control *C. elegans* genomic DNA and yeast genomic DNA, containing Y69A3, are each expected to yield two hybridising bands, while *EcoRI* digests of DNA from the same two sources are each expected to yield a single hybridising band. Initially, the results obtained experimentally do not appear to be in total agreement with the expected results (Figure 3.1) since the *BamHI* digested *C. elegans* genomic DNA (lane 1) generates only a single hybridising band of approximately 23 kb, rather than the two bands expected. However, a closer analysis of the autoradiograph reveals that the perceived single hybridising band in lane 1 actually comprises two bands migrating very close together. Furthermore, the difference seen between the signal intensities of these
two bands in lane 1 is also observed for the yeast DNA in lane 3. Therefore the only discrepancy in the results is due to a reduction in size of one of the Y69A3 hybridising bands from approximately 23 kb to 10 kb. These data suggest that this fragment, and therefore the cpr-3 gene, is located at one end of the cloned insert of Y69A3. If some of this fragment was replaced by YAC vector sequence, as a result of the enzymes used to clone the insert, then a proximal BamHI site in the YAC vector arm could generate the observed decrease in size. The EcoRI digests, in lanes 2 and 4, both produced an identical and much smaller single hybridising band of approximately 4 kb. This suggests that the sequences flanking the gene are still present in the Y69A3 clone, since any replacement with vector sequence must have occurred outwith the EcoRI sites.

3.2.3.2. The hybridisation pattern for cpr-4

The nucleotide sequence of cpr-4 revealed the presence of a single BamHI site near the centre of the coding region of this gene, and two EcoRI sites 670 bp apart. Thus, BamHI digests of control C.elegans genomic DNA and yeast genomic DNA, containing Y7E11, are each expected to yield two hybridising bands. EcoRI digests of DNA from the same two sources are each expected to yield three hybridising bands one of which should be 670 bp in size. The hybridisation pattern predicted from the nucleotide sequence of cpr-4 is also obtained experimentally (Figure 3.2). However, there is a slight discrepancy in the size and hybridisation intensity of the largest band generated by BamHI digestion of control C.elegans DNA (lane 1) and yeast genomic DNA containing Y7E11 (lane 3). Such a discrepancy may be genuine, reflecting the presence of intervening YAC vector sequences, as was observed for cpr-3, or it may be an artefact, since bands of high molecular weight DNA do occasionally migrate aberrantly in digests of genomic DNA and this may be caused by differences in the salt or DNA concentrations between samples. The largest hybridising fragment generated by BamHI digested yeast DNA containing Y7E11 (lane 3) is also fainter than that generated by BamHI digested control C.elegans genomic DNA (lane 1) and this is probably a result of inefficient transfer of the larger DNA fragment to the Hybond N filter during Southern blotting.
3.2.3.3. The hybridisation pattern for cpr-5

The nucleotide sequence of cpr-5 revealed the presence of a single BamHI site near the centre of the coding region of this gene, and one EcoRI site towards the 5' end of the coding region. Thus, BamHI and EcoRI digests of both control C.elegans genomic DNA and yeast genomic DNA, containing Y50D4, are all expected to yield two hybridising bands of unknown size and this hybridisation pattern is obtained experimentally (Figure 3.3A). However, the signal from the 6.6 kb hybridising band of EcoRI digested yeast genomic DNA containing Y50D4 (lane 4) is somewhat less than that of the corresponding band from EcoRI digested control C.elegans genomic DNA (lane 2). This variation may be a result of inefficient transfer of the Y50D4 6.6 kb hybridising fragment to the Hybond N filter during Southern blotting, caused by local variations in transfer efficiency across the surface of the filter.

As well as the expected hybridising bands, strong signals from additional hybridising bands are also obtained with the BamHI digested control C.elegans genomic DNA (lane 1). These additional bands are probably due to slight partial digestion since comparable signals are not obtained with EcoRI digested control C.elegans genomic DNA (lane 2) and are therefore unlikely to be a result of cross hybridisation to other sequences within the C.elegans genome. It should be noted that the same additional hybridising bands are also observed for the BamHI digested C.elegans genomic DNA used as a control for the Southern blot of yeast DNA containing the Y48G5 YAC clone (lane 1, Figure 3.3B). This is not in disagreement with the conclusion that partial digestion may have caused the presence of these additional bands since the BamHI digested C.elegans genomic DNA controls used for the Southern blots of Y50D4 and Y48G5 were obtained from the same digest.

3.2.3.4. The hybridisation pattern for cpr-6

The nucleotide sequence of cpr-6 revealed the presence of a single BamHI site near the centre of the coding region of this gene, and three EcoRI sites. Thus, BamHI digests of control C.elegans genomic DNA and yeast genomic DNA, containing Y55B10 or Y43B9, are each expected to yield two hybridising bands, while EcoRI digests of
DNA from the same two sources are each expected to yield four hybridising bands, two of which are expected to be 380 bp and 1060 bp in size. This predicted hybridisation pattern is obtained experimentally for both Y55B10 and Y43B9 (Figures 3.4A and 4.4B) suggesting that the genomic sequences of *cpr-5* corresponding to the cm01a5 cDNA clone are completely contained within both these YAC clones.

### 3.2.4. The copy number of *cpr-3*, *cpr-4*, *cpr-5* and *cpr-6*

The Southern blots of *BamHI* and *EcoRI* digested *C. elegans* genomic DNA and yeast genomic DNA containing the appropriate YAC clones, probed with the cm12b6, cm14e3, cm04d10 or cm01a5 cDNA clones, indicate that the four cDNA clones hybridise strongly to four distinct loci (Figures 3.1, 3.2, 3.3 and 3.4). Furthermore, the Southern blot analyses produced simple hybridisation patterns for all four genes which could be completely explained by the *BamHI* and *EcoRI* restriction endonuclease sites revealed during later sequencing of *cpr-3*, *cpr-4*, *cpr-5* and *cpr-6*, corresponding to the cDNA clones cm12b6, cm14e3, cm04d10 and cm01a5, respectively. These data strongly suggest that each of these four genes are present as single copies in the genome of *C. elegans* at four distinct loci. From these data alone, it is not possible to discount each of these genes being tandemly repeated at each of their loci. However, later restriction endonuclease mapping of subclones containing each of the four genes strongly suggest this is not the case.

### 3.2.5. *cpr-3*, *cpr-4*, *cpr-5* and *cpr-6* do not cross hybridise significantly

As discussed in Section 3.2.2, the cDNA clones cm12b6, cm14e3, cm04d10 and cm01a5 all generate the greatest signal intensities when hybridising to sequences corresponding to their genomic counterparts, *cpr-3*, *cpr-4*, *cpr-5* and *cpr-6* respectively. These data suggest that the nucleotide sequences of the four genes are sufficiently diverged to prevent significant levels of cross hybridisation of their cDNA clones to each other, or to other cathepsin B-like sequences in the *C. elegans* genome. However, the cDNA clones of *cpr-5* and *cpr-6* (Figures 3.3 and 3.4), do also generate some additional faint signals caused by cross hybridisation to other sequences in the *C. elegans* genome.
This suggests that there may be additional members of the cathepsin B-like multigene family within the genome of *C. elegans*. Alternatively, the additional faint signals may be the result of low levels of cross hybridisation between each of the four genes, *cpr-3*, *cpr-4*, *cpr-5* and *cpr-6*. However, this is unlikely to be the case since later sequence analysis suggest that the nucleotide sequences of these four genes are sufficiently diverged to make cross hybridisation to one another very unlikely (Chapter 4, Section 4.2.5).

The above data also indicate that the hybridisation and wash stringencies used were sufficiently high to distinguish between each of the four genes. These stringencies were therefore used in the experiments described in the subsequent sections of this chapter, in order to isolate genomic clones containing the four genes, *cpr-3*, *cpr-4*, *cpr-5* and *cpr-6*.

3.2.6. Selecting cosmid clones for *cpr-3*, *cpr-4*, *cpr-5* and *cpr-6*

In addition to the overlapping YACs, the *C.elegans* physical map is represented by a set of overlapping cosmid (and lambda) clones with positional correlation between the YAC and cosmid sets (see introduction to this chapter). In order to obtain cloned genomic sequences of each of the four genes, sets of overlapping cosmids were selected on the basis of the positional information gained from the YAC grid physical map data (R. Waterston pers. comm.) and Southern blot data (Figure 3.1, 3.2, 3.3 and 3.4). These clones were then obtained from J. Sulston and A. Coulson at the Sanger Centre.

The physical map data indicate that the *cpr-3* cDNA clone (cm12b6) hybridises to two different YACs, Y46B6 (Figure 3.5A) and Y69A3 (Figure 3.5B), at distinct loci. However, the region covered by Y46B6 is completely overlapped by two additional YAC clones, Y43F3 and Y46C2. Therefore, hybridisation of cm12b6 to Y46B6 should be accompanied by hybridisation to one of these two YAC clones (Figure 3.5A). No such hybridisation was reported and therefore the hybridisation of cm12b6 with Y46B6 on the YAC grid must be artefactual. The Southern blot data obtained using the *cpr-3* cDNA clone (cm12b6) as a probe (Figure 3.1) agreed with this since these data suggest that *cpr-3* is single copy and is contained within the cloned insert of Y69A3. Some regions of the Y69A3 YAC clone are also covered by two additional overlapping YAC clones, Y59A11 and Y45G4 (Figure 3.5B), yet the physical map data suggested that
cm12b6 only hybridised to Y69A3. This suggests that the *cpr-3* gene is contained within the region of the Y69A3 clone not covered by Y59A11 and Y45G4. The four cosmids ZK1037, C29F3, T09D4 and R14F11, which cover this region, were therefore selected and requested from J.Sulston and A.Coulson at the Sanger Centre. However, I was unable to obtain R14F11 because the Sanger Centre experienced problems growing it. Furthermore, there is a region of the Y69A3 YAC clone that is not covered by cosmid clones (indicated in Figure 3.5B). Consequently, I was unable to obtain cosmid clones covering the entire region of Y69A3 expected to contain the *cpr-3* gene.

The physical map data indicated that the *cpr-4* cDNA (cm14e3) hybridises to two overlapping YAC clones, Y7E11 and Y6G12 (Figure 3.6). The Southern blot data obtained using the *cpr-4* cDNA clone (cm14e3) as a probe (Figure 3.2) agree with this since these data suggest that *cpr-4* is single copy and is contained within the cloned insert of Y7E11. Accordingly, the cosmid clones, C40C4, ZK1055, F44C4 and T10H9, which cover the region of overlap of Y7E11 and Y6G12 (Figure 3.6), were selected and requested from J.Sulston and A.Coulson at the Sanger Centre.

The physical map data supplied with cm04d10 were not obtained using this clone but rather the cDNA clone cm5d10. A comparison of the partial sequence data of these two clones reveals that they share 98.7% identity and are therefore assumed to be derived from the same gene, for the reasons discussed in Section 3.2.1. The physical map data indicated that cm5d10 hybridises to two different regions of the YAC grid. This cDNA clone produced a strong hybridising signal within the region covered by Y50D4 (Figure 3.7A) and a weak hybridising signal within the region covered by Y48G5 (Figure 3.7B). The Southern blot analysis obtained using the *cpr-5* cDNA clone (cm04d10) as a probe suggests that *cpr-5* is single copy and contained within the cloned insert of Y50D4 (Figure 3.3A). However, a weak hybridising band was also observed when Southern blots of *EcoRI* digested *C.elegans* genomic DNA and yeast DNA containing the Y48G5 YAC clone were probed with cm04d10 (Figure 3.3B). Together, the physical map and Southern blot data suggest that there may be a gene with some homology to the cm5d10 and cm04d10 cDNA clones in the region covered by the Y48G5 YAC clone. The YAC grid data indicates that cm5d10 hybridises strongly to three overlapping YAC clones, Y50D4, Y50A6 and Y54C3 (Figure 3.7A). Therefore, three cosmids that cover this overlap, W02B2, T21H3 and W06C3, were requested from
J. Sulston and A. Coulson at the Sanger Centre. The YAC grid data also indicates that cm5d10 hybridises weakly to the three overlapping YAC clones Y48G5, Y19H3 and Y41A2 (Figure 3.7B). Therefore, three cosmids that cover this overlap, F09F2, AD10 and F58E8, were also requested from J. Sulston and A. Coulson at the Sanger Centre.

The physical map data indicate that the cpr-6 cDNA clone (cm01a5) hybridises to four overlapping YAC clones; Y43B9, Y55B10, Y44C4 and Y42G7 (Figure 3.8). The Southern blot data obtained using the cpr-6 cDNA clone (cm01a5) as a probe (Figure 3.4A and 3.4B) agreed with this since these data suggest that cpr-6 is single copy and is contained within the cloned inserts of both Y43B9 and K11B10. Therefore, two cosmids, C25B8 and K11B10, covering this region of overlap, were selected and requested from J. Sulston and A. Coulson at the Sanger Centre.

3.2.7. Isolating genomic clones containing cpr-3, cpr-4, cpr-5 and cpr-6

The selected cosmid clones (Section 3.2.6) were screened by colony hybridisation (Chapter 2, Section 2.8.8.6) using the appropriate cDNA clones as probes. For cpr-3, the results indicated that none of the selected cosmid clones hybridised to the cpr-3 cDNA clone (cm12b6). For cpr-4, the two overlapping cosmid clones, ZK1055 and F44C4, were found to hybridise to the cpr-4 cDNA clone (cm14e3). For cpr-5, the W02B2 cosmid clone was found to hybridise to the cpr-5 cDNA clone (cm04d10). The AD10 cosmid also produced a very weak signal with cm04d10, suggesting that there might be a related gene present on this cosmid, subtending the region covered by the YAC Y48G5, as was suggested from the physical map data and the Southern Blot analysis of digested yeast genomic DNA containing Y48G5. For cpr-6, the overlapping cosmid clones, C25B8 and K11B10, were found to hybridise to the cpr-6 cDNA clone (cm01a5). Thus, the cosmid clones; ZK1055 and F44C4 for cpr-4, W02B2 and AD10 for cpr-5 and C25B8 and K11B10 for cpr-6 were selected for further analysis.

From the colony hybridisation, it was apparent that cpr-3 was not contained within the cosmid clones chosen. Therefore, I screened a C. elegans Bristol N2 genomic DNA lambda EMBL4 library obtained from I. Johnstone. Serial dilutions were used to assay the titre of the library, which was found to be 300,000 pfu/ml (Chapter 2, Section
2.9.1). The number of plaques required for the library screen was determined using the formula:

\[ N = \frac{\ln (1-P)}{\ln (1-(1/n))} \]  

(Clark and Carbon, 1976)

Where \( N \) = the number of recombinants, \( P \) = the probability of including any DNA sequence in a random library of \( N \) independent recombinants, \( n \) = the size of the genome relative to a single cloned fragment.

Therefore, assuming: \( P = 0.99 \), \( n = 1 \times 10^8 \) 

\[ \frac{1.5 \times 10^4}{(C.elegans \ genome \ size)} \]  
\[ (average \ insert \ size) \]

\( N = 30,699 \) recombinants

To ensure an even greater probability that the \( cpr-3 \) gene corresponding to \( cm12b6 \) was included, 90,000 plaques were used in the primary screen. This yielded eight putative positive clones. These eight clones were selected and subjected to two further rounds of plaque purification. Seven positives clones named CL#012, CL#034, CL#056, CL#078, CL#0910, CL#01112 and CL#01314 remained after plaque purification and these clones were selected for further analysis as genomic clones of \( cpr-3 \).

3.2.8. Isolation of subclones from the positive cosmid and lambda clones

I decided to subclone the genes from the cosmid and lambda clones into pBluescript II KS- for restriction endonuclease mapping. This approach had the advantage in that the more complex approaches required for mapping cosmid and lambda clones could be avoided. Large fragments which hybridised to the appropriate cDNA clones were subcloned into pBluescript II KS- to reduce the danger of losing sections of the genes, or their flanking sequences, during this subcloning step.
3.2.8.1. Obtaining suitable fragments for subcloning

DNA from the six selected cosmid clones was extracted using the alkaline hydrolysis method and purified by caesium density gradient centrifugation (Chapter 2, Section 2.6.4). The six cosmid clones were digested with a variety of restriction endonucleases, whose recognition sequences were also present in the pBluescript II KS-polylinker, and electrophoresed through 0.8% agarose gels. The gels were Southern blotted, probed with the appropriate cDNA (radioabeled by random priming) and washed at high stringency prior to exposure to X-ray film.

For cpr-4 (Figure 3.9), restriction endonuclease digests of ZK1055 DNA and F44C4 DNA probed with the cpr-4 cDNA clone (cm14e3) generated almost identical hybridisation patterns with the same enzyme. Most of the differences observed were due to partial digestion, confirmed by viewing the ethidium bromide stained agarose gels. Partial digests were observed for; ZK1055 DNA digested with XbaI (lane 2, Figure 3.9A), ZK1055 DNA digested with SacII (lane 6, Figure 3.9A), ZK1055 DNA digested with SpeI (lane 9, Figure 3.9A) and both ZK1055 and F44C4 DNA digested with KpnI (lane 10, Figures 3.9A and 3.9B). The overall similarity in the hybridisation patterns, generated by probing the digested ZK1055 and F44C4 cosmid DNA, suggest that each cosmid carries a copy of the cpr-4 gene, and that the gene is contained entirely within each of the cosmids, assuming that the cpr-4 cDNA clone (cm14e3) used as a probe is nearly full length. However, the partial digest of ZK1055 cosmid DNA with SpeI (lane 9, Figure 3.9A) generates three hybridising bands, as opposed to the single hybridising band resulting from complete digestion of F44C4 cosmid DNA by SpeI (lane 9, Figure 3.9B). The faintest of these three hybridising bands does not appear to correspond to a partial digestion product, suggesting that there may be additional cpr-4 sequences within ZK1055 not present in F44C4. This seems very unlikely since the SacI digests (lane 3, Figures 3.9A and 3.9B) and PstI digests (lanes 5, Figures 3.9A and 3.9B) of ZK1055 and F44C4 both produced single hybridising bands of approximately 4.5 kb and 3.5 kb respectively, for both cosmids, suggesting that cpr-4 was flanked by SacI and PstI sites proximal to this gene. One would not expect such sites to be conserved between the two cosmids if extra sequences from the cpr-4 gene were present in ZK1055 but absent in F44C4. This observation suggests that the anomaly observed was probably an artefact of
partial digestion of ZK1055 by SpeI. However, I decided to use ZK1055 as a source of DNA to isolate subclones containing the \textit{cpr-3} gene, assuming that this cosmid clone might possess additional sequences from this gene.

For \textit{cpr-5} (Figures 3.10A and 10B), only restriction endonuclease digests of W02B2 DNA generated a hybridisation pattern when probed with the \textit{cpr-5} cDNA clone (cm04d10). This suggests that the weak hybridisation of the \textit{cpr-5} cDNA clone (cm04d10) to AD10 observed during colony hybridisation (Section 3.2.7) was probably an artefact, since the wash stringency was the same as that used when probing Southern blots of \textit{EcoRI} digested \textit{C.elegans} genomic DNA and yeast genomic DNA containing Y48G5, which both yielded a weak hybridising band (Figure 3.3B).

For \textit{cpr-6} (Figure 3.11A and 3.11B), only restriction endonuclease digests of C25B8 DNA generated a hybridisation pattern when probed with the \textit{cpr-6} cDNA clone (cm01a5). This indicates that the hybridisation of the \textit{cpr-6} cDNA clone (cm01a5) to K11B10 observed during colony hybridisation (Section 3.2.7) was probably an artefact. Analysis of the ethidium bromide stained agarose gels prior to Southern blotting revealed that partial digestion of C25B8 cosmid DNA had occurred with \textit{XbaI} (lane 2, Figure 3.11A) and \textit{SalI} (lane 4, Figure 3.11A).

The above digests indicate that \textit{cpr-4} is contained within ZK1055 and F44C4, \textit{cpr-5} is contained within W02B2 and \textit{cpr-6} is contained within C25B8. For all three genes, simple hybridisation patterns are obtained, suggesting that these genes are all single copy within the appropriate hybridising clones. For \textit{cpr-4}, I decided to subclone the 6.5 kb hybridising fragment produced by \textit{HindIII} digestion of ZK1055 (lane 8, Figure 3.9A). However, for \textit{cpr-5} and \textit{cpr-6}, none of the single digests of W02B2 and C25B8 generated single hybridising fragments of a useful size for subcloning, these fragments being either too large for efficient subcloning into pBluescript II KS- or too small to aid subcloning of the complete genes including their flanking sequences.

For \textit{cpr-5} and \textit{cpr-6}, double digests were performed using a variety of restriction enzymes which previously produced large (greater than 12 kb) hybridising fragments with single digests of the cosmids W02B2 and C25B8. Double digests were performed using \textit{ApoI}, \textit{SacI} and \textit{PstI}. The digestion products were electrophoresed through 0.7% agarose gels, Southern blotted and probed with the appropriate cDNA as previously described. The results demonstrate that \textit{PstI/SacI} double digests of both W02B2 DNA
and C25B8 DNA produce hybridising fragments of approximately 10 kb useful for subcloning cpr-5 and cpr-6 respectively.

The agarose gels used to identify hybridising fragments suitable for subcloning cpr-4, cpr-5 and cpr-6 from ZK1055, W02B2 and C25B8 could not accurately resolve the size of these fragments. Therefore the digests that were to be used for subcloning DNA fragments containing cpr-4, cpr-5 and cpr-6 from ZK1055, W02B2 and C25B8 were repeated and electrophoresed a greater distance through a 0.7% agarose gel. The agarose gel was Southern blotted, probed with the appropriate cDNA clone and washed at high stringency prior to autoradiography. The results (Figure 3.12) confirm that a 6.8 kb HindIII fragment from cosmid ZK1055 contains the cpr-4 gene, an 11 kb PstI/SacI fragment from cosmid W02B2 contains the cpr-5 gene and a 10.7 kb PstI/SacI fragment from cosmid C25B8 contains the cpr-6 gene.

3.2.8.2. Subcloning fragments containing cpr-4, cpr-5 and cpr-6

A 'shotgun' approach was used to subclone the 6.8 kb HindIII fragment from cosmid ZK1055, the 11 kb PstI/SacI fragment from cosmid W02B2 and the 10.7 kb PstI/SacI fragment from cosmid C25B8. With this approach, the cosmid DNA was digested with the appropriate enzyme(s) and the whole digest used in a ligation reaction with pBluescript II KS- which had been digested with the appropriate enzyme(s) and gel purified. Transformants were then isolated, using ampicillin resistance and blue/white colony selection. Colony hybridisation (Chapter 2, Section 2.8.8.6) was used to screen 50 - 100 transformants for the appropriate insert using the four cDNA clones. The colony hybridisation screen yielded; nine potential positive subclones containing the 6.8 kb HindIII fragment from ZK1055, seven potential positive subclones containing the 11 kb PstI/SacI fragment from W02B2 and-three-potential positive subclones containing the 10.7 kb PstI/SacI fragment from C25B8.

Plasmid DNA was prepared from overnight cultures of the potential positive subclones. The plasmid DNA was digested with the same restriction endonucleases originally used to prepare the ZK1055, W02B2 and C25B8 DNA for subcloning. The digestion products were electrophoresed through 0.7% agarose gels and were visualised by ethidium bromide staining. These digests revealed that five of the nine potential
positive subclones from ZK1055 and one of the three potential positive subclones from C25B8 contained single inserts of approximately the correct size. However, while six of the seven potential positive subclones from W02B2 contained an insert of the correct size, additional inserts were also seen in these six cases. Therefore, I decided to subclone the 11 kb \textit{PstI/SacI} fragment from W02B2 using a more standard approach. One of the six subclones obtained from the colony hybridisation screen was selected as a convenient source of the 11 kb \textit{PstI/SacI} fragment because this insert could be easily isolated from the other contaminating inserts. The 11 kb insert was released from the vector by digestion with \textit{PstI} and \textit{SacI} restriction endonucleases, gel purified and ligated into pBluescript II KS-. Ten colonies were selected after transformation and overnight cultures prepared. The plasmid DNA was extracted, digested with \textit{PstI} and \textit{SacI} restriction endonucleases, electrophoresed through a 0.7% agarose gel, Southern blotted and probed using the \textit{cpr-5} cDNA clone (cm04d10). Analysis of the ethidium bromide stained gel prior to Southern blotting and analysis of the autoradiograph generated after probing revealed that seven of the 10 clones screened carry the desired 11 kb insert from W02B2, and no other contaminating inserts.

Plasmid DNA from the five selected potential positive subclones from ZK1055, and the single potential positive subclone from C25B8, were digested with the appropriate restriction endonucleases. The digests were electrophoresed through a 0.7% agarose gel, Southern blotted and probed with the appropriate cDNA clone. The results from this screen demonstrated that four of the five subclones from ZK1055 possess a single 6.8 kb \textit{HindIII} fragment which hybridised to the \textit{cpr-4} cDNA clone (cm14e3) and thatonesubclone from C25B8 possesses a single 10.7 kb \textit{PstI/SacI} insert which hybridised to the \textit{cpr-6} cDNA clone (cm01a5).

In summary, the genes \textit{cpr-4}, \textit{cpr-5} and \textit{cpr-6}, corresponding to the cDNA clones cm14e3, cm04d10 and cm01a5 respectively, were subcloned into pBluescript II KS- as large (6.8 - 11 kb) fragments. One clone for each gene was selected for further analysis and these were named ZK1055/6.8/KS-, W02B2/11/KS- and C25B8/10.7/KS-, corresponding to the cDNA clones cm14e3, cm04d10 and cm01a5 respectively. Each of the selected subclones contains the gene to which the appropriate cDNA clone hybridises. Digestion with the appropriate restriction endonuclease enzymes and agarose
gel electrophoresis indicate that none of these subclones contain additional, non-contiguous inserts.

3.2.8.3. Isolating the cpr-3 gene from lambda clones

DNA from the seven positive lambda clones CL#012, CL#034, CL#056, CL#078, CL#0910, CL#01112 and CL#01314, obtained from the C.elegans Bristol N2 genomic DNA lambda EMBL4 library screen (Section 3.2.7), was digested with the EcoRI and BamHI restriction endonucleases. The digested DNA was electrophoresed through 0.7% agarose gels, Southern blotted to Hybond N (Amersham International plc) and probed with the cpr-3 cDNA clone (cm12b6).

Figure 3.13A shows the DNA fragments produced after agarose gel electrophoresis of the seven lambda clones digested with EcoRI. EcoRI digestion of all seven lambda clones generates two fragments of 20 kb and 9 kb, representing the left and right arms of the vector respectively. Analysis of the other fragments generated by EcoRI digestion reveals that all seven lambda clones share common bands of approximately 3 kb and 4 kb, suggesting that the inserts of the seven lambda clones all overlap one another. Additional bands unique to individual clones, or groups of clones, were also seen and at least three groups of lambda clones could be distinguished on the basis of larger, easily visible bands. The three groups are represented by CL#012, CL#034 and CL#056. The CL#012 clone appears to be unique, CL#034 may be identical to CL#078 and CL#0910. CL#056 may be identical to CL#01112 and CL#01314. However, smaller bands were obscured by contaminating E.coli genomic DNA making it impossible to determine whether the clones within the latter two groups were truly identical to each other. Figure 3.13B shows the DNA fragments produced after agarose gel electrophoresis of the seven lambda clones digested with BamHI. These results reveal that CL#034 and CL#0910 share a common 2 kb band absent in the clone CL#078. Thus the group of lambda clones comprising CL#034, CL#078 and CL#0910, determined from the EcoRI digests, may be divided further into two groups. Thus the seven lambda clones can be organised into at least four groups based on the pattern of DNA fragments generated by EcoRI and BamHI digests. The autoradiographs obtained after probing the Southern blots of these agarose gels with the
CPR-3 cDNA clone (cm12b6) are shown in Figures 3.14A and 3.14B. The autoradiograph obtained from EcoRI digested DNA from the seven lambda clones (Figure 3.14A) reveals that all these clones contain a common 4 kb hybridising fragment. A 4 kb hybridising fragment was also observed when the CPR-3 cDNA clone (cm12b6) was used to probe Southern blots of EcoRI digested C. elegans Bristol N2 genomic DNA and yeast DNA containing the Y69A3 YAC clone (Figure 3.1). The autoradiographic data obtained from BamHI digested DNA from the seven lambda clones (Figure 3.14B) reveals that CL#034 and CL#0910 share a common 2 kb hybridising band not present in the other lambda clones and therefore one of these two clones, CL#034, was selected for further analysis.

In summary, the seven positive lambda clones analysed are not identical since they can be organised into at least four groups based on the DNA fragment patterns obtained by digestion with EcoRI and BamHI. However, all seven lambda clones possess inserts that overlap one another, including a 4 kb EcoRI fragment that hybridises to the CPR-3 cDNA clone (cm12b6). These data suggest that all seven overlapping lambda clones are derived from a single locus and possess inserts that contain the CPR-3 gene. A single 4 kb hybridising fragment was also observed when Southern blots of EcoRI digested C. elegans Bristol N2 genomic DNA and yeast DNA containing the Y69A3 YAC clone were probed with the cm12b6 cDNA clone of CPR-3 (Figure 3.1). These data suggest that all seven lambda clones are derived from the same locus as that covered by the Y69A3 YAC clone, supporting the conclusion that there is a single copy of the CPR-3 gene within the C. elegans genome. One of these lambda clones, CL#034 was selected for further analysis.

3.2.8.4. Obtaining subclones containing CPR-3 from CL#034

DNA from the lambda clone CL#034 was digested with a variety of restriction endonucleases in both single and double digests. The digests were electrophoresed through 0.7% agarose gels, Southern blotted to Hybond N filters (Amersham International plc) and probed with the CPR-3 cDNA clone (cm12b6). This analysis identified a 5 kb XhoI/SacII fragment that hybridised to cm12b6 (Figure 3.15, lane 6) and this fragment was selected for subcloning. CL#034 DNA was digested with XhoI
and SacII and electrophoresed through a 1.0% agarose gel. The 5 kb XhoI/SacII fragment could not be resolved from two slightly larger fragments. Thus, all three bands were gel purified and ligated into XhoI/SacII digested, gel purified pBluescript II KS+. The ligation mix was transformed into E.coli XL1-blue and positive clones were selected using ampicillin resistance and blue/white colony selection. Plasmid DNA was prepared from overnight cultures of six selected clones. The plasmid DNA was digested with XhoI and SacII, electrophoresed through a 0.7% agarose gel, Southern blotted to Hybond N (Amersham International plc) and probed with the cpr-3 cDNA clone (cm12b6). Analysis of the ethidium bromide stained agarose gel prior to Southern blotting indicated that all six subclones contained single inserts of the correct size. The autoradiographic results indicated that the insert of all six selected subclones hybridised to the cpr-3 cDNA clone. Thus one of the subclones was selected for further analysis and named CL#034/5/KS+.

3.2.9. Restriction endonuclease mapping of the genomic clones

As a first step in restriction mapping, single digests were performed to determine suitable enzymes for mapping each subclone. The digests for each subclone were electrophoresed through 0.6% agarose gels and the bands visualised by ethidium bromide staining. Restriction endonucleases which cut the inserts at three or less sites were selected for restriction mapping.

For restriction mapping, a variety of single and double digests were performed on each of the four subclones. The digested DNA was electrophoresed through 0.7% agarose gels, the DNA fragment patterns generated were visualised by ethidium bromide staining and were recorded on Polaroid film. Southern blots were prepared for each agarose gel using Hybond N (Amersham International plc), probed with the intact cDNA appropriate for each subclone and washed at high stringency. After exposure to medical X-ray film; the probed Southern blots were stripped, reprobed with the 5’ end of each cDNA clone, washed at high stringency and exposed to medical X-ray film. The 5’ end cDNA probes were generated by PCR using the M13 Forward primer and a primer specific to each cDNA clone (Chapter 2, Section 2.13.3) The PCR reaction yielded products of the expected size, 180 bp for cm12b6, 145 bp for cm14e3, 190 bp for
cm04d10 and 230 bp for cm01a5. These PCR products were gel purified and radiolabeled by random priming.

Figures 3.16, 3.17, 3.18 and 3.19 show the results of the digests for each of the four subclones, CL#034/5/KS+, ZK1055/6.8/KS-, W02B2/11/KS- and C25B8/10.7/KS-, obtained by ethidium bromide staining of the agarose gels prior to Southern blotting. Each figure also includes a table that summarises these results, in addition to summarising the results obtained from probing the Southern blots with the intact cDNA and 5’ end cDNA probes. The final restriction maps of the four subclones, CL#034/5/KS+, ZK1055/6.8/KS-, W02B2/11/KS- and C25B8/10.7/KS-, as well as the DNA fragments used to generate these maps, are shown in Figures 3.20, 3.21, 3.22 and 3.23. The position, orientation and maximum length of the four genes cpr-3, cpr-4, cpr-5 and cpr-6 within the appropriate subclone, determined from the Southern blot results using the appropriate intact cDNA and 5’ end cDNA probes, are also shown in these figures.

For all four subclones, the restriction maps were recreated using the smaller fragments generated by the digests because the sizes of these fragments could be determined more accurately. The total insert sizes obtained from the restriction map data for the four subclones are very similar, but not identical, to the insert sizes obtained by Southern blot analysis prior to generating the subclones. Thus, Southern blot analysis predicts the sizes of the DNA fragments used to generate the subclones of cpr-3, cpr-4, cpr-5 and cpr-6 to be 5 kb (Figure 3.15), 6.8 kb (Figure 3.12), 11 kb (Figure 3.12) and 10.7 kb (Figure 3.12), while the restriction map data of these four subclones generates sizes of 4.85 kb, 7 kb, 10.65 kb and 10.5 kb, respectively. The sizes generated from the restriction map data are likely to more accurately reflect the true insert sizes of these four subclones than the Southern blot data, since multiple smaller fragments, whose sizes could be more accurately determined, were used to recreate these restriction maps.

For the two larger subclones, W02B2/11/KS- (Figure 3.22) and C25B8/10.7/KS- (Figure 3.23), some of the restriction sites could not be aligned as accurately as those determined for CL#034/5/KS+ and ZK1055/6.8/KS-, resulting in slight discrepancies in the positions of certain restriction sites. This is assumed to be a result of experimental error, caused by the limits of resolution of the agarose gels used to determine the DNA fragment sizes required for creating the restriction maps. Indeed, discrepancies mostly
arose when restriction sites were predicted from very large fragments, whose sizes could not be determined as accurately. It is not possible to discount the presence of additional restriction sites for the restriction endonucleases used during the mapping experiments which are in close proximity to those sites already determined for these enzymes. Such sites might result in the generation of very small restriction fragments which would not be detectable on the ethidium bromide stained agarose gels. These fragments would necessarily be very small and might therefore also account for the small discrepancies in the insert sizes predicted from the restriction map data and the Southern blot data.

For each of the four subclones, the high degree of similarity in the total insert size, obtained from the restriction map data, to the observed insert size, obtained from Southern blot analysis of the cosmids or lambda clones prior to subcloning, indicates that these subclones do not contain large tandem repeats. Such tandem repeats would result in multiple fragments of the same size and cause a severe underestimation of the insert size predicted by the restriction maps.

3.2.9.1. The position, orientation, maximum size and copy number of cpr-3, cpr-4, cpr-5 and cpr-6 within their subclones

The Southern blot data obtained using the intact cDNA clones as probes revealed the position of each of the four genes within the four subclones, while the Southern blot data obtained using the 5’ end cDNA probes revealed the orientation of each of the four genes within the appropriate subclone. These results are summarised in Figures 3.20, 3.21, 3.22 and 3.23. Unfortunately, these results indicate that the 5’ end of the cpr-3 gene is absent in the CL#034/5/KS+ subclone (Figure 3.20). Thus it was necessary to obtain a second subclone containing the 5’ end of the cpr-3 gene and this is described in Section 3.2.10. Later nucleotide sequencing of the cpr-3 gene revealed a SacII site 116 bp downstream of the start of the putative translation initiation codon. This site was used for subcloning the insert of the CL#034/5/KS+ clone and therefore explains the absence of the 5’ end of the cpr-3 gene within this clone. The nucleotide sequence data also explain why the 5’ end of the cpr-3 gene was not detected in the Southern blot experiments performed prior to subcloning the insert of CL#034/5/KS+ (Figure 3.24). The cDNA clone of cpr-3 (cm12b6) and the 5’ end of the cpr-3 gene share only 111 bp
of common nucleotide sequence upstream of the *SacII* site and this region is interrupted by an intron (Figure 3.24). Thus, there is probably insufficient homology between the 5' ends of the *cpr-3* cDNA clone (cm12b6) and the *cpr-3* gene to produce a hybridising band with the high wash stringencies used.

The restriction map data for each of the four subclones also indicate the maximum size of each gene, assuming that the cDNA clones that were used as probes are nearly full length. This was later demonstrated by sequencing the cDNA clones and conducting 5' Rapid Amplification of cDNA Ends (5' RACE) experiments (Chapter 4, Sections 4.2.2 and 4.2.3). According to the restriction map data, the maximum gene sizes of, *cpr-4*, *cpr-5* and *cpr-6*, contained within the subclones ZK1055/6.8/KS-, W02B2/11/KS- and C25B8/10.7/KS-, are 2.6 kb, 3 kb and 3.7 kb respectively. The maximum length of the portion of the *cpr-3* gene contained within the CL#034/5/KS+ subclone is 1.7 kb. Later sequencing of *cpr-3*, *cpr-4*, *cpr-5* and *cpr-6* revealed their sizes, from the ATG initiation codon to the translation termination codon, to be 1611 bp, 1058 bp, 1269 bp and 1719 bp respectively. For *cpr-3*, 1493 bp of the gene, from the *SacII* site to the translation termination codon, were contained within the CL#034/5/KS+ subclone. These results indicate that the hybridising regions of ZK1055/6.8/KS-, W02B2/11/KS- and C25B8/10.7/KS- are potentially large enough to contain more than one copy of *cpr-4*, *cpr-5* and *cpr-6* respectively. However, this is not the case since the 5' end cDNA probes of these three genes hybridised to DNA fragments from the appropriate subclone which are too small to contain more than one copy of the gene. Thus the 5' end cDNA probes for *cpr-4*, *cpr-5* and *cpr-6* hybridised to DNA fragments of 600 bp, 1400 bp and 700 bp, respectively. Of these fragments, only the 1400 bp *BamHI/XbaI* DNA fragment from W02B2/11/KS-, which hybridised to the 5' end cDNA probe of *cpr-5*, is potentially large enough to contain a second copy of this gene. However, later sequencing of the *cpr-5* gene revealed that over 700 bp of this gene is contained within this fragment, leaving insufficient room for a second copy. Thus all four subclones contain only a single copy of the appropriate gene as expected from previous data which suggest that each of these genes are present as single copies within the genome of *C. elegans*. 

105
3.2.9.2. The nucleotide sequences of cpr-3, cpr-4, cpr-5 and cpr-6 support the restriction mapping data

Sequencing of the cpr-3, cpr-4, cpr-5 and cpr-6 (Chapter 4) revealed restriction endonuclease sites within these four genes that were in complete agreement with the results obtained from the restriction mapping experiments and these results are summarised in Figures 3.25, 3.26, 3.27 and 3.28. It should be noted that these figures contain additional information about subclones generated later. As discussed in Section 3.2.9.1, the nucleotide sequence of cpr-3 explained why the 5' end of this gene was truncated in the CL#034/5/KS+ clone and also explained why the 5' end of this gene was not detected by the Southern blot experiments performed prior to subcloning the insert of CL#034/5/KS+.

The restriction map of CL#034/5/KS+ (summarised in Figure 3.25) predicts the presence of a SacII site, two HindIII sites, a ClaI site and a PstI site in that order within the cpr-3 gene. Nucleotide sequencing of the cpr-3 gene confirmed the presence of these restriction endonuclease sites in the same order as predicted from the restriction map. Furthermore, the nucleotide sequence of cpr-3 reveals that these sites are separated from one another by approximately the same number of nucleotides as expected from the restriction map. Thus, nucleotide sequencing indicates that the first HindIII site is 259 bp downstream of the SacII site, the second HindIII site is 911 bp downstream of the first HindIII site, the ClaI site is 350 bp downstream of the second HindIII site and the PstI site is 194 bp downstream of the ClaI site, and the restriction map data predicted these sites to be separated by 250 bp, 900 bp, 400 bp and 150 bp, respectively. In addition, the restriction map data suggest that no XbaI, EcoRI or XhoI restriction endonuclease sites are present in the cpr-3 gene and this is also confirmed by the nucleotide sequence of cpr-3.

The restriction map of ZK1055/6.8/KS- (summarised in Figure 3.26) predicted the presence of a single BamHI site within the cpr-4 gene and a single KpnI site approximately 600 bp upstream of the BamHI site, possibly within the 5' flank of cpr-4. The nucleotide sequence of cpr-4 reveals that this is the case, by identifying a single BamHI site within the cpr-4 gene (370 bp downstream of the start of the ATG translation initiation codon) and a single KpnI site 516 bp upstream of the BamHI site (in
the 5' flank of this gene). In addition, the restriction map data suggest that no *PstI*, *XbaI*, *HindIII* or *ClaI* restriction endonuclease sites are present in the *cpr-4* gene and this is also confirmed by the nucleotide sequence of *cpr-4*.

The restriction map of W02B2/11/KS- (summarised in Figure 3.27) predicted the presence of a single *BamHI* site within the *cpr-5* gene and this is confirmed by the nucleotide sequence of *cpr-5*, which identifies a single *BamHI* site 711 bp downstream of the start of the ATG translation initiation codon. In addition, the restriction map data suggest that no *XbaI*, *XhoI*, *SacI*, *PstI* or *ClaI* restriction endonuclease sites are present in the *cpr-5* gene and this is also confirmed by the nucleotide sequence of *cpr-5*.

The restriction map of C25B8/10.7/KS- (summarised in Figure 3.28) predicted the presence of a single *BamHI* site within the *cpr-6* gene and a single *SalI* site approximately 700 bp upstream of the *BamHI* site, possibly within the 5' flank of this gene. The nucleotide sequence of *cpr-6* revealed that this is the case, by identifying a single *BamHI* site within the *cpr-6* gene (725 bp downstream of the start of the ATG translation initiation codon) and a single *SalI* site 734 bp upstream of the *BamHI* site (9 bp upstream of the start of the ATG translation initiation codon). In addition, the restriction map data suggest that no *SacI*, *PstI* or *KpnI* restriction endonuclease sites are present in the *cpr-6* gene and this is also confirmed by the nucleotide sequence of *cpr-6*.

In summary, the restriction maps generated for the four subclones, CL#034/5/KS+, ZK1055/6.8/KS-, W02B2/11/KS- and C25B8/10.7/KS- agree with the nucleotide sequence obtained for *cpr-3*, *cpr-4*, *cpr-5* and *cpr-6* contained within these subclones. These data provide solid evidence for: the efficacy of each of the restriction maps, the single copy nature of each gene within the appropriate subclone and the correct positioning and orientation of each gene within the appropriate subclone.

3.2.10. Subcloning the 5' flank of *cpr-3* from CL#034

The restriction map data of CL#034/5/KS+ (Figure 3.20) was used to select enzymes which might be useful for subcloning the 5' end of the *cpr-3* gene. Mapped sites for *HindIII*, *ClaI*, *XbaI*, *PstI* and *EcoRI* sites were available in the subclone CL#034/5/KS+. Therefore, the lambda clone CL#034 was digested with each of these enzymes, the products electrophoresed through a 0.8% agarose gel and Southern blotted
to Hybond N (Amersham International plc). The blot was then probed with the 250 bp 
*HindIII/SacII* fragment from CL#034/5/KS+. This fragment contained part of the 5’ end 
of *cpr-3* present within CL#034/5/KS+. The results of this (Figure 3.29) revealed that 
*HindIII, ClaI* and *EcoRI* generated fragments upstream of, but overlapping, the 
CL#034/5/KS+ clone. The 1.7 kb *HindIII* and 3 kb *ClaI* fragments were gel purified, 
subcloned into pBluescript II KS+ and transformed into *E.coli* XL1-blue. Plasmid DNA 
was extracted from overnight cultures of selected transformants. The cloned 1.7 kb 
*HindIII* fragment was digested with *HindIII* while the cloned 3 kb *ClaI* fragment was 
digested with both *ClaI* and *HindIII*. These data indicated that both fragments had been 
subcloned and therefore a clone from each was selected and named CL#034/1.7/KS+ and 
CL#034/3/KS+ respectively. The restriction map data of CL#034/5/KS+ suggested that 
both subclones CL#034/1.7/KS+ and CL#034/3/KS+ would contain approximately the 
same amount of 5’ flank sequence from the *cpr-3* gene. The *ClaI/HindIII* double digest 
of CL#034/3/KS+ indicated that it only contained approximately 75 bp more 5’ flank 
sequence than the 1.7 kb *HindIII* fragment of CL#034/1.7/KS+, and this was confirmed 
by sequencing into the 5’ end of CL#034/3/KS+ using the T3 primer. Though 
CL#034/3/KS+ provided a small amount of extra 5’ flank sequence useful for later *lacZ* 
fusion constructs, the 1.7 kb *HindIII* fragment could be more readily subcloned into the 
lacZ fusion vectors to be used, which contained a *HindIII* site in their polylinkers. Thus 
CL#034/1.7/KS+ was selected for later use in the generation lacZ fusion constructs for 
cpr-3.

3.2.11. The generation of subclones from ZK1055/6.8/KS-, W02B2/11/KS- and 
C25B8/10.7/KS-

Subclones containing the 5’ flanks of the three genes *cpr-4, cpr-5* and *cpr-6* were 
generated from the subclones ZK1055/6.8/KS-, W02B2/11/KS- and C25B8/10.7/KS- 
respectively, for later use in the generation of *lacZ* fusion constructs. The restriction 
and 3.23) were used to determine the appropriate fragments for subcloning into 
pBluescript II KS+. The 2.8 kb *BamHI/HindIII* fragment of ZK1055/6.8/KS-, the 3 kb 
*BamHI* fragment of W02B2/11/KS-, the 3.7 kb *BamHI/PstI* fragment of
C25B8/10.7/KS- and the 3 kb *BamHI/SalI* fragment of C25B8/10.7/KS- were gel purified and ligated into pBluescript II KS+. The ligation products were transformed into *E.coli* XL1-blue and several transformants were selected for each of the subclones prepared. Plasmid DNA was prepared from overnight cultures of these transformants, digested with the appropriate restriction endonucleases and electrophoresed through 0.7% agarose gels to check the sizes of the inserts. One of each of the subclones containing the 2.8 kb *BamHI/HindIII* fragment of ZK1055/6.8/KS-, the 3.7 kb *BamHI/PstI* fragment of C25B8/10.7/KS- and the 3 kb *BamHI/SalI* fragment of C25B8/10.7/KS- were selected for further analysis and named ZK1055/2.8/KS+, C25B8/3.7/KS+ and C25B8/3/KS+. Since the 3 kb *BamHI* insert could ligate into the pBluescript II KS+ vector in one of two orientations, several clones containing inserts of the correct size were selected for further analysis. The orientation of the 3 kb *BamHI* insert of W02B2/3/KS+ in these clones was determined by sequencing into the ends using the T7 and KS primers. Two clones were selected, each containing the insert in the opposite orientation, and were named W02B2/3/KS+#1 and W02B2/3/KS+#2.

The subclone W02B2/11/KS+ proved to be unstable in *E.coli* XL1-blue cells. The subclone was therefore transformed into *E.coli* SURE competent cells (Stratagene Ltd.) which are deficient in most of the *E.coli* recombination pathways. The subclone W02B2/11/KS+ proved to be slightly more stable in *E.coli* SURE cells, however deleted clones were still frequently isolated after extracting DNA from overnight cultures of these cells. In contrast the clone W02B2/3/KS+ showed no such instability in *E.coli* XL1-blue cells. Therefore, I decided to subclone a smaller fragment from W02B2/11/KS+, containing the 3′ end of the *cpr-5* gene. The restriction map of W02B2/11/KS+ (Figure 3.22) was used to determine the appropriate fragment to clone. The 1.6 kb *PstI/BamHI* fragment was chosen, gel purified and ligated into the pBluescript II KS+ vector. The ligation products were transformed into *E.coli* XL1-blue cells and several transformants were selected. Plasmid DNA was prepared from overnight cultures of these transformants, digested with the appropriate restriction endonucleases and electrophoresed through 0.7% agarose gels to check the sizes of the inserts. One clone which contained the correct size of insert was selected for further analysis and named W02B2/1.6/KS+. The cloning of the 1.6 kb *PstI/BamHI* fragment from W02B2/11/KS+ was not ideal since this fragment was adjacent to but did not
overlap W02B2/3/KS+. Therefore, it was possible that some sequences might be lost if two BamHI sites were present in close proximity. However, later sequence comparison of the cDNA clone cm04d10 with the cpr-5 gene, which spans both subclones, revealed this is not the case. The relationships of the subclones generated for each of the four genes, including the restriction enzymes sites used for cloning and the positions of the appropriate genes (determined by restriction mapping and sequencing), are summarised in Figures 3.25, 3.26, 3.27 and 3.28.

3.2.12. Determining whether the subclones are contiguous with the *C. elegans* genome

It was important to assess whether or not the subclones generated were contiguous with the *C. elegans* genome, particularly for those subclones required for generating lacZ fusion constructs to study the tissue specific expression patterns of cpr-3, cpr-4, cpr-5 and cpr-6 (see Chapter 5). Two screens were performed to assess whether the subclones containing cpr-3, cpr-4, cpr-5 and cpr-6 were contiguous with the *C. elegans* genome. In the first screen, the subcloned inserts were analysed to determine whether they contained fragments from the cosmid and lambda vectors used to generate the original clones. In the second screen, diagnostic digests were performed with *C. elegans* Bristol N2 genomic DNA using restriction endonucleases whose recognition sites had been mapped within the various subclones. This analysis was performed to determine whether the DNA fragment sizes and fragment patterns that were expected from the subclones were generated by the digested *C. elegans* genomic DNA after hybridisation with the appropriate cDNA clone.

For the first screen, the subclones containing cpr-3, cpr-4, cpr-5 and cpr-6 were digested with the appropriate restriction endonucleases to release the subcloned inserts. The digests were electrophoresed through 1.0% agarose gels, Southern blotted to Hybond N (Amersham International plc) and hybridised with sequences homologous to the original cosmid and lambda vectors. The results of these experiments demonstrated that none of the inserts from the subclones containing cpr-3, cpr-4 or cpr-5 hybridised with the vector probes, indicating that the inserts from these subclones do not possess any of the original cosmid or lambda vector sequences. However, the insert of subclone
C25B8/10.7/KS-, containing the \textit{cpr-6} gene, did produce a weak hybridising signal when probed with pBluescript sequences, in addition to the strong hybridising signal expected from the approximately 3 kb pBluescript II KS- vector of this subclone (Figure 3.30). The cosmid vector pJB8 was used to generate the cosmid clone C25B8 from which C25B8/10.7/KS- was generated. Since the pJB8 and pBluescript vectors are both derived from the pBR322 plasmid, the hybridisation of pBluescript sequences to the cloned 10.7 kb insert of C25B8/10.7/KS- suggests that some original pJB8 cosmid vector sequence is present in this insert. The cosmid pJB8 possesses a \textit{PstI} site approximately 700 bp from the \textit{BamHI} site which was originally used for cloning the inserts obtained by \textit{Sau3A} partial digestion of \textit{C.elegans} genomic DNA during preparation of the cosmid library. The insert of C25B8/10.7/KS- was cloned using \textit{PstI} and \textit{SacI} sites. Therefore, the results suggest that there are 700 bp of pJB8 sequence present in this subclone and in the subclones C25B8/3.7/KS+ and C25B8/3/KS+, which also contain inserts cloned via the same \textit{PstI} site (Figure 3.28). To confirm this, a single sequencing reaction was performed using the T3 primer and C25B8/3/KS+. The nucleotide sequence generated shares complete base identity to sequences adjacent to the \textit{PstI} site of pJB8, within the 700 bp region of this vector defined by the \textit{PstI} and \textit{BamHI} sites. These results indicate that C25B8/3/KS+, C25B8/3.7/KS+ and C25B8/10.7/KS-, whose inserts were all cloned via the same \textit{PstI} site, possess approximately 700 bp of the pJB8 cosmid vector within their inserts.

Since the pJB8 cosmid vector sequence is upstream of the mapped position of the \textit{cpr-6} gene, it was necessary to determine how much 5' flank sequence of the \textit{cpr-6} gene was contained within the three subclones C25B8/3/KS+, C25B8/3.7/KS+ and C25B8/10.7/KS- downstream of the pJB8 vector sequence. In order to do this, new restriction endonuclease sites within the 5' flank of \textit{cpr-6}, between the gene and the pJB8 vector sequences, had to be determined to provide sites for diagnostic digests of \textit{C.elegans} genomic DNA. Later sequencing of \textit{cpr-6}, revealed several \textit{HindIII} sites present within the gene. These data were used with the restriction map data to generate a partial restriction map for C25B8/3.7/KS+ (Figure 3.31A). \textit{HindIII} single and \textit{HindIII/SalI} double digests were used to determine the position of the \textit{HindIII} site upstream of the \textit{SalI} site (Figure 3.31B). These results indicate that there is a \textit{HindIII} site 1.35 kb upstream of the \textit{SalI} site in this subclone as well as in C25B8/3/KS+ and
C25B8/10.7/KS-, which cover the same region of the *C. elegans* genome. This site was later used for diagnostic digests of *C. elegans* genomic DNA.

In the second screen, *C. elegans* genomic DNA was digested with restriction endonucleases, whose sites had been mapped in the subclones containing the four genes. The DNA was electrophoresed through 0.7% agarose gels, Southern blotted to Hybond N (Amersham International plc) and hybridised with the cDNA clone appropriate for the set of digests. In some cases, the sites recognised by the enzymes used had not been mapped to all the subclones for a given gene. In these instances, the subclones were also digested with the same restriction endonucleases, Southern blotted to Hybond N (Amersham International plc) and probed with the appropriate cDNA clone to allow comparison with the pattern of hybridising fragments produced by the digested *C. elegans* genomic DNA. The restriction endonucleases were chosen after consulting the restriction map data for each of the subclones summarised in Figures 3.25, 3.26, 3.27 and 3.28.

Figure 3.32 shows the results obtained for the diagnostic digests used to determine whether the CL#034/5/KS+ and CL#034/1.7/KS+ subclones, containing the *cpr-3* gene are contiguous with the *C. elegans* genome. The *EcoRV* restriction endonuclease was used in some of the diagnostic digests. Though an *EcoRV* site had been mapped in CL#034/1.7/KS+, it was unknown whether additional *EcoRV* sites were present in CL#034/5/KS+. Therefore, a triple digest was performed with CL#034/5/KS+ DNA using the *EcoRV*, *SacII* and *HindIII* enzymes. The result obtained after agarose gel electrophoresis of the digest, Southern blotting and probing with the *cpr-3* cDNA clone (cm12b6) is shown in Figure 3.32A. The hybridisation pattern observed indicates that no *EcoRV* sites are present in this subclone, since the hybridisation pattern observed is consistent with cleavage of the insert DNA at the mapped *HindIII* and *SacII* restriction sites and release of the remaining insert DNA from the vector by cleavage at the *EcoRV* site within the pBluescript II KS+ polylinker. The expected hybridisation patterns, resulting from triple digests of *C. elegans* genomic DNA using *EcoRV* as one of the enzymes, are based on this observation (Figure 3.32C). The hybridisation patterns generated by the diagnostic digests of *C. elegans* genomic DNA are shown in Figure 3.32B while the expected hybridisation patterns are shown Figure 3.32C, based on the restriction map data obtained for CL#034/5/KS+ and CL#034/1.7/KS+ (summarised in
Figure 3.25). The hybridisation patterns obtained from the diagnostic digests were entirely as expected and therefore indicate that the CL#034/5/KS+ and CL#034/1.7/KS+ subclones are contiguous with the *C.elegans* genome.

Figure 3.33A shows the results obtained for the diagnostic digests used to determine whether the ZK1055/6.8/KS- subclone containing the *cpr-4* gene is contiguous with the *C.elegans* genome. The expected hybridisation patterns for the various digests are also summarised (Figure 3.33B), based on the restriction map data for ZK1055/6.8/KS- (summarised in Figure 3.26). The hybridisation patterns obtained from the diagnostic digests are entirely as expected from the restriction map data and therefore indicate that the ZK1055/6.8/KS- subclone is entirely contiguous with the *C.elegans* genome.

Figure 3.34A shows the hybridisation patterns obtained for the diagnostic digests used to determine whether the W02B2/3/KS+ and W02B2/1.6/KS+ subclones, containing the *cpr-5* gene, are contiguous with the *C.elegans* genome. The larger W02B2/11/KS- subclone was not analysed because this clone was unstable and therefore all subsequent experiments were performed using the two smaller subclones. Figure 3.34B shows the expected hybridisation patterns for the digests shown in Figure 3.34A. These expected patterns were obtained from the restriction map data for W02B2/3/KS+ and W02B2/1.6/KS+ (summarised in Figure 3.27). Again, the hybridisation patterns observed from the diagnostic digests are as expected from the restriction map data for these two subclones and therefore indicate that W02B2/3/KS+ and W02B2/1.6/KS+ are contiguous with the *C.elegans* genome.

For the subclones containing the *cpr-6* gene (C25B8/10.7/KS-, C25B8/3.7/KS+ and C25B8/3.0/KS+), a slightly different approach was used. These clones are not entirely contiguous with the *C.elegans* genome as a result of the presence of pJB8 cosmid vector sequence within the inserts (discussed above). Therefore, the *HindIII* restriction endonuclease was used in diagnostic digests of *C.elegans* genomic DNA because an internal *HindIII* site, downstream of the pJB8 sequence had been mapped in the C25B8/3.7/KS- subclone (discussed above). However, it was not known whether additional *HindIII* sites might be present in the region of the insert of C25B8/10.7/KS- not covered by the smaller C25B8/3.7/KS- subclone and outwith the region sequenced (Figure 3.28). Thus, both the largest subclone (C25B8/10.7/KS-) and *C.elegans*
genomic DNA were digested using single and double digests involving *HindIII*. The results obtained after agarose gel electrophoresis of the digests, Southern blotting and probing with the *cpr-6* cDNA clone (cm01a5) are shown in Figure 3.35A and Figure 3.35B, along with the expected fragment patterns (Figure 3.35C), based on the known *HindIII* sites from the restriction map and sequence data for C25B8/10.7/KS-. The hybridisation patterns from the digested *C.elegans* genomic DNA and the digested C25B8/10.7/KS- DNA are identical. Furthermore, the fragment sizes and fragment patterns are as expected from the *HindIII* sites determined by restriction mapping or sequencing (summarised in Figure 3.28) with the exception of the absence of a 0.15 kb hybridising fragment for *HindIII/SalI* double digests of both *C.elegans* genomic DNA and C25B8/10.7/KS- DNA, which was due to this small fragment running off the end of the gel. However, an additional hybridising band of approximately 1.3 kb was also observed for *C.elegans* genomic DNA and C25B8/10.7/KS- DNA with *HindIII* single or *HindIII/SalI* double digests. Analysis of the ethidium bromide stained agarose gel of the C25B8/10.7/KS- DNA digests prior to Southern blotting suggest the additional 1.3 kb hybridising band is a result of partial digestion. This conclusion is supported by previous Southern blots of C25B8 cosmid DNA digested with *HindIII* and hybridised with the *cpr-6* cDNA clone (cm01a5) which only generated the four expected bands (lane 8, Figure 3.11A). Furthermore, *HindIII/BamHI* double digests of both C25B8/10.7/KS- DNA and *C.elegans* genomic DNA only produce the expected bands (lane 2 Figure 3.35A and 3.35B).

The results of these experiments indicate that a previously undetermined *HindIII* site is present 2 kb downstream of the most 3' *HindIII* site identified from sequencing the *cpr-6* gene. This site is approximately 1.65 kb downstream of the putative translation termination codon of *cpr-6*. The identical hybridisation patterns observed between the digested subclone and the digested *C.elegans* genomic DNA indicate that the region of C25B8/10.7/KS- covered by these *HindIII* sites is contiguous with the *C.elegans* genome. Importantly, this indicates that the region of the 5' flank of *cpr-6*, up to the mapped *HindIII* site 1.35 kb upstream of the *SalI*, site is contiguous with the *C.elegans* genome and that this contiguity extends to approximately 1.65 kb downstream of the translation termination codon of *cpr-6*. 

114
3.2.13. The physical map positions of \textit{cpr-3}, \textit{cpr-4}, \textit{cpr-5} and \textit{cpr-6}

As stated in the discussion to this chapter, there is good alignment between the physical and genetic maps of \textit{C. elegans}. Thus, information regarding the position of genes obtained from the physical map of \textit{C. elegans} can be related to the genetic map. The physical map positions of the three cosmid clones ZK1055, W02B2 and C25B8, containing \textit{cpr-4}, \textit{cpr-5} and \textit{cpr-6} respectively, were already known prior to isolation of the genes from these cosmids. However, no physical map data were available for the CL\#034 lambda clone and therefore this clone was sent to J.Sulston and A.Coulson at the Sanger Centre for physical mapping. According to their analysis, CL\#034 (and therefore \textit{cpr-3}) was found to map to the Y69A3 YAC clone as expected from the physical map data for the \textit{cpr-3} cDNA (cm12b6). Another cathepsin B-like gene, \textit{cpr-1}, previously isolated from \textit{C. elegans} (Ray and McKerrow, 1992) also maps to this region. Restriction endonuclease fingerprinting performed by A.Coulson indicated that CL\#034 almost entirely overlaps the lambda clone from which \textit{cpr-1} was isolated, suggesting that these clones contain two distinct but very closely linked genes, or that \textit{cpr-1} and \textit{cpr-3} are the same gene. Later sequencing of the \textit{cpr-3} gene confirmed that \textit{cpr-1} and \textit{cpr-3} are distinct genes, having different gene architectures and coding for proteins with only 57.4\% amino acid identity (Chapter 4, Sections 4.2.5 and 4.2.7). Alignment of the physical and genetic maps reveal that \textit{cpr-3} maps to the right end of the central region of chromosome V, between the \textit{unc-76} gene (map position +7.02) and the stP18 restriction fragment length polymorphism (RFLP, map position +9.69).

Alignment of the physical and genetic maps reveals that \textit{cpr-4} and \textit{cpr-5} are both on chromosome V but are not closely linked to one another or to \textit{cpr-1} and \textit{cpr-3}. The \textit{cpr-4} gene, contained in the ZK1055 cosmid clone, maps to the centre of chromosome V between the stP192 RFLP (map position +0.01) and the \textit{snb-1} gene (map position +0.18). The \textit{cpr-5} gene, contained in the W02B2 cosmid clone, maps to the left end of chromosome V between the two RFLPs, nP61 (map position -20.23) and nP62 (map position -19.54).

In contrast, alignment of the physical and genetic maps reveals that \textit{cpr-6}, contained in the C25B8 cosmid clone, is not linked to \textit{cpr-1}, \textit{cpr-3}, \textit{cpr-4} or \textit{cpr-5}. This
gene maps to the centre of the X chromosome between *sup-7* (map position -2.9) and *unc-6* (map position -2.18).

### 3.3. Discussion

The experiments described in this chapter had two purposes; to determine whether the genome of *C. elegans* encodes a multigene family of cathepsin B-like enzymes and to clone such genes for future characterisation and analysis. The four cathepsin B-like genes, *cpr-3*, *cpr-4*, *cpr-5* and *cpr-6*, were cloned using four previously isolated cDNA clones (Waterston et al., 1992). Later sequencing of the four genes (reported in Chapter 4) demonstrated that these genes are cathepsin B-like. The three genes, *cpr-4*, *cpr-5* and *cpr-6*, were cloned using the physical map of *C. elegans* and the physical map data supplied with the cDNA clones. This approach enabled selected sets of cosmid clones to be screened on the basis of the positional information obtained from the physical map data. These screens resulted in the isolation of 3 cosmid clones, ZK1055, W02B2 and C25B8, which contained the 3 genes *cpr-4*, *cpr-5* and *cpr-6* respectively. The *cpr-3* gene was isolated by screening a *C. elegans* genomic DNA lambda EMBL4 library which yielded the clone CL#034, containing this gene. The four genes were subcloned and these subclones were subsequently restriction mapped.

The isolation of four cathepsin B-like genes from *C. elegans* indicates that this free living nematode species possesses a cathepsin B-like multigene family with at least five members, comprising the four genes reported here and the previously isolated *cpr-1* gene (Ray and McKerrow, 1992). The physical map data and results from Southern blot analyses of digested *C. elegans* DNA and yeast genomic DNA, containing the appropriate YAC clone, indicate that each of the four genes, *cpr-3*, *cpr-4*, *cpr-5* and *cpr-6*, are located at a distinct locus within the *C. elegans* genome. These Southern Blot analyses also indicate that the four genes are sufficiently diverged not to cross hybridise strongly to one another or to other sequences in the *C. elegans* genome. The restriction map data for the subclones containing these genes indicate that each of the genes is single copy within its respective subclone or subclones. Together, these data indicate that these four genes are single copy, are diverged from one another and are dispersed throughout the genome of *C. elegans* and therefore suggest that they have arisen from an ancient gene.
duplication event. The dispersed genome locations of these four genes may have enabled them to evolve independently since these genes would no longer be subjected to the homogenising influences normally associated with tandem arrays, such as gene conversion and unequal crossover events.

The approximate positions of the clones containing the four genes, cpr-3, cpr-4, cpr-5 and cpr-6, within the genome were determined by alignment of the physical and genetic maps of C.elegans. Four of the genes are located on chromosome V but only two genes are closely linked, the cpr-3 gene and the previously isolated cpr-1 gene. Despite their close linkage, these genes do not appear to be tandem repeats since later sequencing indicates that these two genes are quite diverged, possessing different gene architectures and encoding proteins with only 57.4% identity. The other two genes present on chromosome V, cpr-4 and cpr-5, are not closely linked to one another or to cpr-1 and cpr-3 and therefore clearly do not represent tandem repeats. The cpr-6 gene, is not linked to the other four genes, being located on the X chromosome.

In addition to allowing the approximate location of the four genes to be determined, alignment of the physical and genetic maps also yields interesting candidate mutant phenotypes in the region of the genetic map covering the physical map location of these genes. For cpr-4, the let-473, let-462 and let-473 lethal mutations were identified which have early or early / mid larval lethal phenotypes. For cpr-5, the let-326 lethal mutation which has an early larval lethal phenotype was identified. Both these genes have temporal expression patterns consistent with these mutant phenotypes since expression of both genes is induced during late embryogenesis and is elevated during the larval stages of development (Chapter 5). For cpr-6, the flu-2 mutation was identified. This mutation results in reduced gut fluorescence and may be a result of defects in the catabolic pathways which occur in the gut. The cpr-6 gene has a tissue specific expression pattern consistent with this mutation, since this gene appears to be expressed exclusively in the intestinal cells of C.elegans (Chapter 5). In order to determine whether these mutant phenotypes are indeed a result of mutations in the cpr-4, cpr-5 or cpr-6 genes, plasmid rescue experiments could be performed. With such experiments, C.elegans strains carrying the desired mutation are transformed with plasmids containing wild-type copies of the gene to be tested. The progeny are then screened to determine whether the plasmid is able to rescue the mutant phenotype.
Figure 3.1

Southern blot analysis of *C. elegans* genomic DNA and yeast genomic DNA, containing the YAC clone Y69A3, both digested with BamHI or EcoRI

The digested *C. elegans* genomic DNA (5µg) and digested yeast genomic DNA (1µg) were electrophoresed through a 0.8% agarose gel, Southern blotted to Hybond N (Amersham International plc), hybridised with the *cpr-3* cDNA clone (cm12b6) and washed at high stringency prior to exposure to medical X-ray film.

**Lane 1:** *C. elegans* genomic DNA digested with BamHI

**Lane 2:** *C. elegans* genomic DNA digested with EcoRI

**Lane 3:** yeast genomic DNA containing Y69A3 digested with BamHI

**Lane 4:** yeast genomic DNA containing Y69A3 digested with EcoRI
Figure 3.2

Southern blot analysis of *C. elegans* genomic DNA and yeast genomic DNA, containing the YAC clone Y7E11, both digested with *BamHI* or *EcoRI*.

The digested *C. elegans* genomic DNA (5 μg) and digested yeast genomic DNA (1 μg) were electrophoresed through a 0.8% agarose gel, Southern blotted to Hybond N (Amersham International plc), hybridised with the *cpr-4* cDNA clone (cm14e3) and washed at high stringency prior to exposure to medical X-ray film.

**Lane 1:** *C. elegans* genomic DNA digested with *BamHI*  
**Lane 2:** *C. elegans* genomic DNA digested with *EcoRI*  
**Lane 3:** yeast genomic DNA containing Y7E11 digested with *BamHI*  
**Lane 4:** yeast genomic DNA containing Y7E11 digested with *EcoRI*
Figure 3.3

Southern blot analysis of BamHI or EcoRI digested C. elegans genomic DNA and yeast genomic DNA, containing the YAC clone Y50D4 (A) or Y48G5 (B)

The digested C. elegans genomic DNA (5μg) and digested yeast genomic DNA (1μg) were electrophoresed through a 0.8% agarose gel, Southern blotted to Hybond N (Amersham International plc), hybridised with the cpr-5 cDNA clone (cm04d10) and washed at high stringency prior to exposure to medical X-ray film.

A. Lane 1: C. elegans genomic DNA digested with BamHI
     Lane 2: C. elegans genomic DNA digested with EcoRI
     Lane 3: yeast genomic DNA containing Y50D4 digested with BamHI
     Lane 4: yeast genomic DNA containing Y50D4 digested with EcoRI

B. Lane 1: C. elegans genomic DNA digested with BamHI
     Lane 2: C. elegans genomic DNA digested with EcoRI
     Lane 3: yeast genomic DNA containing Y48G5 digested with BamHI
     Lane 4: yeast genomic DNA containing Y48G5 digested with EcoRI
Southern blot analysis of *BamHI* or *EcoRI* digested *C. elegans* genomic DNA and yeast genomic DNA, containing the YAC clone Y55B10 (A) or Y43B9 (B).

The digested *C. elegans* genomic DNA (5µg) and digested yeast genomic DNA (1µg) were electrophoresed through a 0.8% agarose gel, Southern blotted to Hybond N (Amersham International plc), hybridised with the *cpr-6* cDNA clone (cm01a5) and washed at high stringency prior to exposure to medical X-ray film.

**A.**
- Lane 1: *C. elegans* genomic DNA digested with *BamHI*
- Lane 2: *C. elegans* genomic DNA digested with *EcoRI*
- Lane 3: yeast genomic DNA containing Y55B10 digested with *BamHI*
- Lane 4: yeast genomic DNA containing Y55B10 digested with *EcoRI*

**B.**
- Lane 1: *C. elegans* genomic DNA digested with *BamHI*
- Lane 2: *C. elegans* genomic DNA digested with *EcoRI*
- Lane 3: yeast genomic DNA containing Y43B9 digested with *BamHI*
- Lane 4: yeast genomic DNA containing Y43B9 digested with *EcoRI*
Figure 3.5

A summary of the physical map data for the cpr-3 cDNA clone (cm12b6) supplied by R. Waterston

The figure shows the YAC clones covering the two regions of the YAC grid to which cm12b6 hybridised.

A. The region of the YAC grid covered by Y46B6.
B. The region of the YAC grid covered by Y69A3.

A and B do not show all the physical map data available for each region. They only show the data that were used to select cosmid clones which might contain the cpr-3 gene and the cosmid clones that were selected. A and B were redrawn from information obtained from ACEDB.
Figure 3.6

A summary of the physical map data for the *cpr*-4 cDNA clone (cm14e3) supplied by R. Waterston.

The figure shows the YAC clones covering the single region of the YAC grid to which cm14e3 hybridised. The figure does not show all the physical map data available for each region. It only shows the data that were used to select cosmid clones which might contain the *cpr*-4 gene and the cosmid clones that were selected. The figure was redrawn from information obtained from ACEDB.
KEY:

- YAC clones
- cosmid clones
- strong hybridising signal
- weak hybridising signal
Figure 3.7

A summary of the physical map data for the cm5d10 cDNA clone supplied by R. Waterston

This cDNA clone is assumed to be derived from the cpr-5 gene because partial sequencing revealed that cm5d10 shares 98.7% nucleotide sequence identity with the cm04d10 cDNA clone used to isolate this gene. The figure shows the YAC clones covering the two regions of the YAC grid to which cm5d10 hybridised.

A. The region of the YAC grid covered by Y50D4.
B. The region of the YAC grid covered by Y48G5.

A and B do not show all the physical map data available for each region. They only show the data that were used to select cosmid clones which might contain the cpr-3 gene and the cosmid clones that were selected. A and B were redrawn from information obtained from ACEDB.
**A**

- Y50A6
- **Y50D4**
- **Y54C3**
- W06C3
- T21H3
- W02B2

**B**

- *Y19H3*
- *Y41A2*  
- *Y48G5*
- F09F2
- F58E8
- **AD10**

**KEY:**

- **YAC clones**
- **cosmid clones**
- **strong hybridising signal**
- *weak hybridising signal*
Figure 3.8

A summary of the physical map data for the cpr-6 cDNA clone (cm01a5) supplied by R. Waterston

The figure shows the YAC clones covering the single region of the YAC grid to which cm01a5 hybridised. The figure does not show all the physical map data available for each region. It only shows the data that were used to select cosmid clones which might contain the cpr-6 gene and the cosmid clones that were selected. The figure was redrawn from information obtained from ACEDB.
The results of Southern blot analysis of ZK1055 and F44C4 cosmids DNA.

Cosmid DNA was digested with a variety of restriction endonucleases. The digests were electrophoresed through 0.8% agarose gels, Southern blotted to Hybond N (Amersham International plc) and hybridised with the *cpr-4* cDNA clone (cm14e3). The filters were washed at high stringency prior to exposure with Medical X-ray film.

### A. ZK1055 DNA digests

<table>
<thead>
<tr>
<th>Lane 1</th>
<th>Lane 2</th>
<th>Lane 3</th>
<th>Lane 4</th>
<th>Lane 5</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>ApaI</em></td>
<td><em>XbaI</em></td>
<td><em>SacI</em></td>
<td><em>SalI</em></td>
<td><em>PstI</em></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Lane 6</th>
<th>Lane 7</th>
<th>Lane 8</th>
<th>Lane 9</th>
<th>Lane 10</th>
<th>Lane 11</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>SacII</em></td>
<td><em>XhoI</em></td>
<td><em>HindIII</em></td>
<td><em>SpeI</em></td>
<td><em>KpnI</em></td>
<td><em>ClaI</em></td>
</tr>
</tbody>
</table>

### B. F44C4 DNA digests

<table>
<thead>
<tr>
<th>Lane 1</th>
<th>Lane 2</th>
<th>Lane 3</th>
<th>Lane 4</th>
<th>Lane 5</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>ApaI</em></td>
<td><em>XbaI</em></td>
<td><em>SacI</em></td>
<td><em>SalI</em></td>
<td><em>PstI</em></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Lane 6</th>
<th>Lane 7</th>
<th>Lane 8</th>
<th>Lane 9</th>
<th>Lane 10</th>
<th>Lane 11</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>SacII</em></td>
<td><em>XhoI</em></td>
<td><em>HindIII</em></td>
<td><em>SpeI</em></td>
<td><em>KpnI</em></td>
<td><em>ClaI</em></td>
</tr>
</tbody>
</table>
The results of Southern blot analysis of W02B2 and AD10 cosmid DNA

Cosmid DNA was digested with a variety of restriction endonucleases. The digests were electrophoresed through 0.8% agarose gels, Southern blotted to Hybond N (Amersham International plc) and hybridised with the cpr-5 cDNA clone (cm04d10). The filters were washed at high stringency prior to exposure with Medical X-ray film.

A. W02B2 DNA digests

<table>
<thead>
<tr>
<th>Lane 1</th>
<th>Lane 2</th>
<th>Lane 3</th>
<th>Lane 4</th>
<th>Lane 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>ApaI</td>
<td>XbaI</td>
<td>SacI</td>
<td>SalI</td>
<td>PslI</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Lane 6</th>
<th>Lane 7</th>
<th>Lane 8</th>
<th>Lane 9</th>
<th>Lane 10</th>
<th>Lane 11</th>
</tr>
</thead>
<tbody>
<tr>
<td>SacII</td>
<td>XhoI</td>
<td>HindIII</td>
<td>SpeI</td>
<td>KpnI</td>
<td>ClaI</td>
</tr>
</tbody>
</table>

B. AD10 DNA digests

<table>
<thead>
<tr>
<th>Lane 1</th>
<th>Lane 2</th>
<th>Lane 3</th>
<th>Lane 4</th>
<th>Lane 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>ApaI</td>
<td>XbaI</td>
<td>SacI</td>
<td>SalI</td>
<td>PslI</td>
</tr>
</tbody>
</table>
The results of Southern blot analysis of C25B8 and K11B10 cosmid DNA

Cosmid DNA was digested with a variety of restriction endonucleases. The digests were electrophoresed through 0.8% agarose gels, Southern blotted to Hybond N (Amersham International plc) and hybridised with the cpr-6 cDNA clone (cm01a5). The filters were washed at high stringency prior to exposure with Medical X-ray film.

A. C25B8 DNA digests

<table>
<thead>
<tr>
<th>Lane 1</th>
<th>Lane 2</th>
<th>Lane 3</th>
<th>Lane 4</th>
<th>Lane 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>ApaI</td>
<td>XbaI</td>
<td>SacI</td>
<td>SalI</td>
<td>PstI</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Lane 6</th>
<th>Lane 7</th>
<th>Lane 8</th>
<th>Lane 9</th>
<th>Lane 10</th>
<th>Lane 11</th>
</tr>
</thead>
<tbody>
<tr>
<td>SacII</td>
<td>XhoI</td>
<td>HindIII</td>
<td>SpeI</td>
<td>KpnI</td>
<td>ClaI</td>
</tr>
</tbody>
</table>

B. K11B10 DNA digests

<table>
<thead>
<tr>
<th>Lane 1</th>
<th>Lane 2</th>
<th>Lane 3</th>
<th>Lane 4</th>
<th>Lane 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>ApaI</td>
<td>XbaI</td>
<td>SacI</td>
<td>SalI</td>
<td>PstI</td>
</tr>
</tbody>
</table>
Southern blots of restriction endonuclease digested ZK1055, W02B2 and C25B8 cosmid DNA

The restriction endonuclease digested cosmid DNA was electrophoresed through a 0.7% gel and Southern blotted to Hybond N (Amersham International plc). The blots were hybridised with the appropriate cDNA clone and washed at high stringency prior to exposure to medical X-ray film.

**Lane 1:** ZK1055 DNA digested with *Hind*III and probed with the *cpr*-4 cDNA clone (cm14e3)

**Lane 2:** W02B2 DNA digested with *Pst*I and *Sac*I and probed with the *cpr*-5 cDNA clone (cm04d10)

**Lane 3:** C25B8 DNA digested with *Pst*I and *Sac*I and probed with the *cpr*-6 cDNA clone (cm01a5)
Figure 3.13

Ethidium bromide stained 0.7% agarose gels of restriction endonuclease digested DNA from the seven positive lambda clones; CL#012, CL#034, CL#056, CL#078, CL#0910, CL#01112 and CL#01314

A. *EcoRI* digests.

<table>
<thead>
<tr>
<th>Lane 1</th>
<th>Lane 2</th>
<th>Lane 3</th>
<th>Lane 4</th>
<th>Lane 5</th>
<th>Lane 6</th>
<th>Lane 7</th>
</tr>
</thead>
<tbody>
<tr>
<td>CL#012</td>
<td>CL#034</td>
<td>CL#056</td>
<td>CL#078</td>
<td>CL#0910</td>
<td>CL#01112</td>
<td>CL#01314</td>
</tr>
</tbody>
</table>

B. *BamHI* digests.

<table>
<thead>
<tr>
<th>Lane 1</th>
<th>Lane 2</th>
<th>Lane 3</th>
<th>Lane 4</th>
<th>Lane 5</th>
<th>Lane 6</th>
<th>Lane 7</th>
</tr>
</thead>
<tbody>
<tr>
<td>CL#012</td>
<td>CL#034</td>
<td>CL#056</td>
<td>CL#078</td>
<td>CL#0910</td>
<td>CL#01112</td>
<td>CL#01314</td>
</tr>
</tbody>
</table>
Figure 3.14

Southern blots of restriction endonuclease digested DNA from the seven positive lambda clones; CL#012, CL#034, CL#056, CL#078, CL#0910, CL#01112 and CL#01314

The Southern blots were hybridised with the \textit{cpr}-3 cDNA clone (cm12b6) and washed at high stringency prior to exposure to medical X-ray film.

A. \textit{EcoRI} digests.

<table>
<thead>
<tr>
<th>Lane 1</th>
<th>Lane 2</th>
<th>Lane 3</th>
<th>Lane 4</th>
<th>Lane 5</th>
<th>Lane 6</th>
<th>Lane 7</th>
</tr>
</thead>
<tbody>
<tr>
<td>CL#012</td>
<td>CL#034</td>
<td>CL#056</td>
<td>CL#078</td>
<td>CL#0910</td>
<td>CL#01112</td>
<td>CL#01314</td>
</tr>
</tbody>
</table>

B. \textit{BamHI} digests.

<table>
<thead>
<tr>
<th>Lane 1</th>
<th>Lane 2</th>
<th>Lane 3</th>
<th>Lane 4</th>
<th>Lane 5</th>
<th>Lane 6</th>
<th>Lane 7</th>
</tr>
</thead>
<tbody>
<tr>
<td>CL#012</td>
<td>CL#034</td>
<td>CL#056</td>
<td>CL#078</td>
<td>CL#0910</td>
<td>CL#01112</td>
<td>CL#01314</td>
</tr>
</tbody>
</table>
Figure 3.15

Southern blot of CL.034 DNA digested with a variety of restriction endonucleases and electrophoresed through a 0.7% agarose gel

The Southern blot was hybridised with the cpr-3 cDNA clone (cm12b6) and washed at high stringency prior to exposure to medical X-ray film.

Lane 1: PstI/XbaI
Lane 2: PstI/SacII
Lane 3: PstI/XhoI
Lane 4: XbaI/SacII
Lane 5: XbaI/XhoI
Lane 6: XhoI/SacII
Lane 7: ApaI/KpnI
Figure 3.16

Restriction map data for CL#034/5/KS+

**A.** Ethidium bromide stained 0.7% agarose gels of CL#034/5/KS+ DNA digested with a variety of restriction endonucleases

<table>
<thead>
<tr>
<th>Lane 1:</th>
<th>Lane 2:</th>
<th>Lane 3:</th>
<th>Lane 4:</th>
<th>Lane 5:</th>
<th>Lane 6:</th>
<th>Lane 7:</th>
<th>Lane 8:</th>
<th>Lane 9:</th>
<th>Lane 10:</th>
<th>Lane 11:</th>
</tr>
</thead>
<tbody>
<tr>
<td>HindIII</td>
<td>HindIII/SacI</td>
<td>HindIII/XhoI</td>
<td>PstI/SacI</td>
<td>EcoR1/SacI</td>
<td>EcoR1/XhoI</td>
<td>XbaI/SacI</td>
<td>XbaI/XhoI</td>
<td>ClaI/XhoI</td>
<td>ClaI</td>
<td>XbaI</td>
</tr>
</tbody>
</table>

**B.** Table summarising the DNA fragment sizes generated by the digests shown in **A**, which were used to generate the restriction map of CL#034/5/KS+. The table also summarises the results obtained after hybridising Southern blots of the gels shown in **A** with the intact cDNA (cm12b6) and the 5' end cDNA of cpr-3.
<table>
<thead>
<tr>
<th>Lane Number</th>
<th>Enzymes</th>
<th>Fragment Sizes</th>
<th>Fragment Hybridising to cDNA clone</th>
<th>Fragment Hybridising to 5' end of cDNA clone</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lane 1</td>
<td>HindIII</td>
<td>7 kb, 0.9 kb</td>
<td>7 kb, 0.9 kb</td>
<td>7 kb</td>
</tr>
<tr>
<td>Lane 2</td>
<td>HindIII / SacII</td>
<td>6.8 kb, 0.9 kb, 0.25 kb</td>
<td>6.8 kb, 0.9 kb, 0.25 kb (faint)</td>
<td>0.25 kb</td>
</tr>
<tr>
<td>Lane 3</td>
<td>HindIII / Xhol</td>
<td>3.7 kb, 3.1 kb, 0.9 kb</td>
<td>3.7 kb, 3.1 kb, 0.9 kb</td>
<td>3.1 kb</td>
</tr>
<tr>
<td>Lane 4</td>
<td>PstI / SacII</td>
<td>6 kb, 1.7 kb</td>
<td>1.7 kb</td>
<td>1.7 kb</td>
</tr>
<tr>
<td>Lane 5</td>
<td>EcoRI / SacII</td>
<td>Partial Digest</td>
<td>No Data</td>
<td>No Data</td>
</tr>
<tr>
<td>Lane 6</td>
<td>EcoRI / Xhol</td>
<td>6 kb, 1.7 kb</td>
<td>6 kb</td>
<td>6 kb</td>
</tr>
<tr>
<td>Lane 7</td>
<td>XbaI / SacII</td>
<td>4.5 kb, 2.7 kb, 0.5 kb</td>
<td>2.7 kb</td>
<td>2.7 kb</td>
</tr>
<tr>
<td>Lane 8</td>
<td>XbaI / Xhol</td>
<td>5.8 kb, 1.5 kb, 0.5 kb</td>
<td>5.8 kb</td>
<td>5.8 kb</td>
</tr>
<tr>
<td>Lane 9</td>
<td>CiaI / Xhol</td>
<td>4.6 kb, 2.2 kb, 1.1 kb</td>
<td>4.5 kb</td>
<td>4.5 kb</td>
</tr>
<tr>
<td>Lane 10</td>
<td>CiaI</td>
<td>5.8 kb, 2.2 kb</td>
<td>No Data</td>
<td>No Data</td>
</tr>
<tr>
<td>Lane 11</td>
<td>XbaI</td>
<td>7.5 kb, 0.5 kb</td>
<td>No Data</td>
<td>No Data</td>
</tr>
</tbody>
</table>
Figure 3.17

Restriction map data for ZK1055/6.8/KS-

A. Ethidium bromide stained 0.7% agarose gels of ZK1055/6.8/KS- DNA digested with a variety of restriction endonucleases

Lane 1: *BamH*I  
Lane 2: *Clal*  
Lane 3: *BamH*I/*Kpn*I  
Lane 4: *Kpn*I  
Lane 5: *BamH*I/*Cla*I  
Lane 6: *Pst*I  
Lane 7: *Pst*I/*Cla*I  
Lane 8: *Pst*I/*BamH*I  
Lane 9: *Pst*I/*Kpn*I  
Lane 10: *XbaI*  
Lane 11: *XbaI*/*BamH*I  
Lane 12: *XbaI*/*Cla*I  
Lane 13: *XbaI*/*Pst*I  
Lane 14: *XbaI*/*Kpn*I

B. Table summarising the DNA fragment sizes generated by the digests shown in A, which were used to generate the restriction map of ZK1055/6.8/KS-. The table also summarises the results obtained after hybridising Southern blots of the gels shown in A with the intact cDNA (cm14c3) and the 5’ end cDNA of *cpr*-4.
<table>
<thead>
<tr>
<th>Lane Number</th>
<th>Enzymes</th>
<th>Fragment Sizes</th>
<th>Fragment Hybridising to cDNA clone</th>
<th>Fragment Hybridising to 5' end of cDNA clone</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lane 1</td>
<td>BamHI</td>
<td>7.2 kb, 2.8 kb</td>
<td>7.2 kb, 2.8 kb</td>
<td>2.8 kb</td>
</tr>
<tr>
<td>Lane 2</td>
<td>ClaI</td>
<td>5.7 kb, 4.2 kb</td>
<td>5.7 kb</td>
<td>5.7 kb</td>
</tr>
<tr>
<td>Lane 3</td>
<td>BamHI / KpnI</td>
<td>4.1 kb, 2.9 kb, 2.2 kb, 0.6 kb</td>
<td>4.1 kb, 0.6 kb</td>
<td>0.6 kb</td>
</tr>
<tr>
<td>Lane 4</td>
<td>KpnI</td>
<td>5.3 kb, 4.6 kb</td>
<td>4.6 kb</td>
<td>4.6 kb</td>
</tr>
<tr>
<td>Lane 5</td>
<td>BamHI / ClaI</td>
<td>4.1 kb, 2.9 kb, 1.55 kb, 1.2 kb</td>
<td>4.1 kb, 1.55 kb</td>
<td>1.55 kb</td>
</tr>
<tr>
<td>Lane 6</td>
<td>PstI</td>
<td>5.1 kb, 3.3 kb, 1.5 kb</td>
<td>3.3 kb</td>
<td>3.3 kb</td>
</tr>
<tr>
<td>Lane 7</td>
<td>PstI / ClaI</td>
<td>PstI Digest Failed</td>
<td>No Data</td>
<td>No Data</td>
</tr>
<tr>
<td>Lane 8</td>
<td>PstI / BamHI</td>
<td>5.1 kb, 2.0 kb, 1.5 kb, 1.3 kb</td>
<td>2.0 kb, 1.3 kb</td>
<td>1.3 kb</td>
</tr>
<tr>
<td>Lane 9</td>
<td>PstI / KpnI</td>
<td>2.9 kb, 2.5 kb, 2.1 kb, 1.5 kb, 0.8 kb</td>
<td>2.5 kb</td>
<td>2.5 kb</td>
</tr>
<tr>
<td>Lane 10</td>
<td>Xbal</td>
<td>4.9 kb, 3.7 kb, 1.2 kb</td>
<td>3.7 kb</td>
<td>3.7 kb</td>
</tr>
<tr>
<td>Lane 11</td>
<td>Xbal / BamHI</td>
<td>4.9 kb, 2.1 kb, 1.55 kb, 1.2 kb</td>
<td>2.1 kb, 1.55 kb</td>
<td>1.55 kb</td>
</tr>
<tr>
<td>Lane 12</td>
<td>Xbal / ClaI</td>
<td>3.7 kb, 2.9 kb, 2.0 kb, 1.2 kb</td>
<td>3.7 kb</td>
<td>3.7 kb</td>
</tr>
<tr>
<td>Lane 13</td>
<td>Xbal / PstI</td>
<td>4.9 kb, 3.3 kb, 1.2 kb, (0.3 kb)*</td>
<td>3.3 kb</td>
<td>3.3 kb</td>
</tr>
<tr>
<td>Lane 14</td>
<td>Xbal / KpnI</td>
<td>2.9 kb, 2.7 kb, 2.0 kb, 1.2 kb, 1.0 kb</td>
<td>2.7 kb</td>
<td>2.7 kb</td>
</tr>
</tbody>
</table>

* predicted band, not visualised
Figure 3.18

Restriction map data for W02B2/11/KS-

A. Ethidium bromide stained 0.7% agarose gels of W02B2/11/KS- DNA digested with a variety of restriction endonucleases

Lane 1: XhoI
Lane 2: XhoI/PstI
Lane 3: XhoI/XbaI
Lane 4: ClaI
Lane 5: ClaI/SacI
Lane 6: BamHI
Lane 7: BamHI/PstI
Lane 8: BamHI/SacI
Lane 9: BamHI/ClaI

Lane 10: XhoI/SacI
Lane 11: XhoI/BamHI
Lane 12: XhoI/ClaI
Lane 13: XbaI
Lane 14: XbaI/SacI
Lane 15: XbaI/PstI
Lane 16: XbaI/ClaI
Lane 17: XbaI/BamHI

B. Table summarising the DNA fragment sizes generated by the digests shown in A, which were used to generate the restriction map of W02B2/11/KS-. The table also summarises the results obtained after hybridising Southern blots of the gels shown in A with the intact cDNA (cm04d10) and the 5’ end cDNA of cpr-5.
<table>
<thead>
<tr>
<th>Lane Number</th>
<th>Enzymes</th>
<th>Fragment Sizes</th>
<th>Fragment Hybridising to Intact cDNA</th>
<th>Fragment Hybridising to 5' end of cDNA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lane 1</td>
<td>Xhol</td>
<td>6.2 kb, 6.2 kb, 1.1 kb</td>
<td>6.2 kb</td>
<td>6.2 kb</td>
</tr>
<tr>
<td>Lane 2</td>
<td>Xhol / PstI</td>
<td>6.2 kb, 6.2 kb, 1.1 kb</td>
<td>6.2 kb</td>
<td>6.2 kb</td>
</tr>
<tr>
<td>Lane 3</td>
<td>Xhol / XbaI</td>
<td>4.2 kb, 3.1 kb, 3.1 kb, 2.2 kb, 0.9 kb, (0.3 kb)*</td>
<td>3.1 kb</td>
<td>3.1 kb</td>
</tr>
<tr>
<td>Lane 4</td>
<td>ClaI</td>
<td>8 kb, 5.7 kb</td>
<td>5.7 kb</td>
<td>5.7 kb</td>
</tr>
<tr>
<td>Lane 5</td>
<td>ClaI / SacI</td>
<td>5.7 kb, 4.9 kb, 2.9 kb</td>
<td>5.7 kb</td>
<td>5.7 kb</td>
</tr>
<tr>
<td>Lane 6</td>
<td>BamHII</td>
<td>10.5 kb, 3.0 kb</td>
<td>10.5 kb, 3.0 kb</td>
<td>3.0 kb</td>
</tr>
<tr>
<td>Lane 7</td>
<td>BamHII / PstI</td>
<td>9 kb, 3.0 kb, 1.6 kb</td>
<td>3.0 kb, 1.6 kb</td>
<td>3.0 kb</td>
</tr>
<tr>
<td>Lane 8</td>
<td>BamHII / SacI</td>
<td>Digest Failed</td>
<td>No data</td>
<td>No data</td>
</tr>
<tr>
<td>Lane 9</td>
<td>BamHII / ClaI</td>
<td>7.8 kb, 3.0 kb, 1.6 kb, 1.0 kb</td>
<td>3.0 kb, 1.6 kb</td>
<td>3.0 kb</td>
</tr>
<tr>
<td>Lane 10</td>
<td>Xhol / SacI</td>
<td>6.2 kb, 3.6 kb, 2.9 kb, 1.1 kb</td>
<td>6.2 kb</td>
<td>6.2 kb</td>
</tr>
<tr>
<td>Lane 11</td>
<td>Xhol / BamHII</td>
<td>6.2 kb, 3.0 kb, 1.6 kb, 1.4 kb, 1.1 kb</td>
<td>3 kb, 1.6 kb</td>
<td>3.0 kb</td>
</tr>
<tr>
<td>Lane 12</td>
<td>Xhol / ClaI</td>
<td>6.2 kb, 5.8 kb, 1.1 kb, (0.4 kb)*</td>
<td>5.8 kb</td>
<td>5.8 kb</td>
</tr>
<tr>
<td>Lane 13</td>
<td>XbaI</td>
<td>7.2 kb, 4 kb, 2.5 kb</td>
<td>7.2 kb</td>
<td>7.2 kb</td>
</tr>
<tr>
<td>Lane 14</td>
<td>XbaI / SacI</td>
<td>5.8 kb, 4 kb, 2.5 kb, 1.3 kb</td>
<td>5.8 kb</td>
<td>5.8 kb</td>
</tr>
<tr>
<td>Lane 15</td>
<td>XbaI / PstI</td>
<td>4.2 kb, 4 kb, 3 kb, 2.5 kb</td>
<td>3 kb</td>
<td>3 kb</td>
</tr>
<tr>
<td>Lane 16</td>
<td>XbaI / ClaI</td>
<td>4.2 kb, 3 kb, 2.7 kb, 2.5 kb, 1.25 kb</td>
<td>3 kb</td>
<td>3 kb</td>
</tr>
<tr>
<td>Lane 17</td>
<td>XbaI / BamHII</td>
<td>5.9 kb, 2.5 kb, 2.4 kb, 1.65 kb, 1.4 kb</td>
<td>5.9 kb, 1.4 kb</td>
<td>1.4 kb</td>
</tr>
</tbody>
</table>

* denote predicted bands (could not visualise on gel)
Restriction map data for C25B8/10.7/KS-

A. Ethidium bromide stained 0.7% agarose gels of C25B8/10.7/KS- DNA digested with a variety of restriction endonucleases

<table>
<thead>
<tr>
<th>Lane 1:</th>
<th>KpnI</th>
<th>Lanes 7 and 11:</th>
<th>BamHI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lane 2:</td>
<td>KpnI/PstI</td>
<td>Lane 8:</td>
<td>BamHI/KpnI</td>
</tr>
<tr>
<td>Lane 3:</td>
<td>KpnI/SacI</td>
<td>Lane 9:</td>
<td>BamHI/PstI</td>
</tr>
<tr>
<td>Lane 4:</td>
<td>SalI</td>
<td>Lane 10:</td>
<td>BamHI/SacI</td>
</tr>
<tr>
<td>Lane 5:</td>
<td>SalI/SacI</td>
<td>Lane 12:</td>
<td>BamHI/SalI</td>
</tr>
<tr>
<td>Lane 6:</td>
<td>SalI/KpnI</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

B. Table summarising the DNA fragment sizes generated by the digests shown in A, which were used to generate the restriction map of C25B8/10.7/KS-. The table also summarises the results obtained after hybridising Southern blots of the gels shown in A with the intact cDNA (cm01a5) and the 5' end cDNA of cpr-6.
<table>
<thead>
<tr>
<th>Lane Number</th>
<th>Enzymes</th>
<th>Fragment Sizes</th>
<th>Fragment Hybridising to intact cDNA clone</th>
<th>Fragment Hybridising to 5' end of cDNA clone</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lane 1</td>
<td>KpnI</td>
<td>9.5 kb, 4.2 kb</td>
<td>9.5 kb</td>
<td>9.5 kb</td>
</tr>
<tr>
<td>Lane 2</td>
<td>KpnI / PstI</td>
<td>9.5 kb, 4.2 kb</td>
<td>9.5 kb</td>
<td>9.5 kb</td>
</tr>
<tr>
<td>Lane 3</td>
<td>KpnI / ScaI</td>
<td>9.5 kb, 2.9 kb, 1.4 kb</td>
<td>9.5 kb</td>
<td>9.5 kb</td>
</tr>
<tr>
<td>Lane 4</td>
<td>SalI</td>
<td>11 kb, 3 kb</td>
<td>11 kb</td>
<td>11 kb</td>
</tr>
<tr>
<td>Lane 5</td>
<td>SalI / ScaI</td>
<td>7.8 kb, 3 kb, 2.9 kb</td>
<td>7.8 kb</td>
<td>7.8 kb</td>
</tr>
<tr>
<td>Lane 6</td>
<td>SalI / KpnI</td>
<td>6.5 kb, 4.2 kb, 3 kb</td>
<td>6.5 kb</td>
<td>6.5 kb</td>
</tr>
<tr>
<td>Lane 7 and 11</td>
<td>BamHI</td>
<td>11 kb, 3 kb</td>
<td>11 kb (faint), 3 kb</td>
<td>11 kb</td>
</tr>
<tr>
<td>Lane 8</td>
<td>BamHI / KpnI</td>
<td>4.2 kb, 3.8 kb, 3 kb, 2.4 kb</td>
<td>3 kb, 3.8 kb (faint)</td>
<td>3.8 kb</td>
</tr>
<tr>
<td>Lane 9</td>
<td>BamHI / PstI</td>
<td>7 kb, 3.8 kb, 3 kb</td>
<td>3.8 kb</td>
<td>3.8 kb</td>
</tr>
<tr>
<td>Lane 10</td>
<td>BamHI / ScaI</td>
<td>6.9 kb, 3.9 kb, 3 kb</td>
<td>6.9 kb</td>
<td>6.9 kb</td>
</tr>
<tr>
<td>Lane 12</td>
<td>BamHI / SalI</td>
<td>7 kb, 3 kb, 3 kb, 0.7 kb</td>
<td>3 kb, 0.7 kb</td>
<td>0.7 kb</td>
</tr>
</tbody>
</table>
Figure 3.20

The restriction map of CL#034/5/KS+

The restriction map shows the position of each mapped restriction endonuclease site and the distances between these sites (in kb). The map also shows the position, orientation (5' - 3') and maximum size of the cpr-3 gene within the subclone (marked by the arrow), obtained from the hybridisation data. The data used to generate the restriction map of CL#034/5/KS+ (shown in Figure 3.16A and 3.16B) are also summarised below the map.
Figure 3.21

The restriction map of ZK1055/6.8/KS-

The restriction map shows the position of each mapped restriction endonuclease site and the distances between these sites (in kb). The map also shows the position, orientation (5' - 3') and maximum size of the cpr-4 gene within the subclone (marked by the arrow), obtained from the hybridisation data. The data used to generate the restriction map of ZK1055/6.8/KS- (shown in Figure 3.17A and 3.17B) are also summarised below the map.

Note: The position of the HindIII cloning site with respect to the restriction endonuclease sites in the polylinker of pBluescript II KS- is indicated by a vertical line on the appropriate fragments.
Position and orientation of the gene

Restriction endonuclease sites within the polylinker

KEY:

B= BamHI
C= ClaI
K= KpnI
P= PstI
X= XbaI

* = cloning site

- Indicates fragment that hybridises to the 5' end cDNA probe
- Indicates fragment that hybridises to the intact cDNA probe

Total Size: 7 kb
Scale: 0.5 kb
The restriction map of W02B2/11/KS-

The restriction map shows the position of each mapped restriction endonuclease site and the distances between these sites (in kb). The map also shows the position, orientation (5’ - 3’) and maximum size of the cpr-5 gene within the subclone (marked by the arrow), obtained from the hybridisation data. The data used to generate the restriction map of W02B2/11/KS- (shown in Figure 3.18A and 3.18B) are also summarised below the map.

Note: The position of the PsrI and SacI cloning site with respect to the restriction endonuclease sites in the polylinker of pBluescript II KS- are indicated by a vertical line on the appropriate fragments.
The restriction map of C25B8/10.7/KS-

The restriction map shows the position of each mapped restriction endonuclease site and the distances between these sites (in kb). The map also shows the position, orientation (5' - 3') and maximum size of the cpr-6 gene within the subclone (marked by the arrow), obtained from the hybridisation data. The data used to generate the restriction map of C25B8/10.7/KS- (shown in Figure 3.19A and 3.19B) are also summarised below the map.

Note: The position of the PstI and SacI cloning site with respect to the restriction endonuclease sites in the polylinker of pBluescript II KS- are indicated by a vertical line on the appropriate fragments.
Figure 3.24

A summary of the nucleotide sequence identity shared between *cpr-3* and its cDNA clone, cm12b6, upstream of the *SacII* site used for cloning the insert of CL#034/5/KS+
**gagttttcagttttccctcaaatttca**

-80 tgtgcggtgttgttttgctgaacgctttctctctcagttttccctcaaatttca

**aaaattgataactaaagagaATGCTGAAAAGTGACTTTTTTG**

-20 aaaattgataactaaagagaATGCTGAAAAGTGACTTTTTTGgtgagcatctttcatcttc

**MLKIVFL**

41 **GCACTTTTCTAGCCGGGCTCTGACATTG**

**ALFLAGCSAFV**

101 **TTCTTGATGAAAATCCGC**

**SacII**

**TTCTTGATGAAAATCCGC**

**SacII**

**KEY:**

**Bold:**  Nucleotide sequence of the 5' end of the cpr-3 cDNA clone (cm12b6) up to and including the SacII site.

**Underlined:**  Nucleotide sequence of the 5' end of the cpr-3 gene up to and including the SacII site.

**Uppercase:**  Coding sequence, the amino acid residue encoded by each codon is given below the appropriate codon.

**Lowercase:**  Non-coding sequence
A summary of the restriction map and DNA sequence data obtained from the subclones, CL#034/5/KS+ and CL#034/1.7/KS+

The figure shows the position and size of the two subclones, CL#034/5/KS+ and CL#034/1.7/KS+ with respect to one another. The position of the restriction endonuclease sites determined for each of these subclones by restriction mapping are also shown. A summary of the DNA sequence of the $cpr{-}3$ gene, obtained from the two subclones, is shown above the subclones. This summary shows the orientation of the gene (5'-3', marked by the arrow), the coding regions of the gene (marked by black boxes), the introns (marked by the thin lines between boxes) and the mapped restriction endonuclease sites that were confirmed by sequencing. Above the DNA sequence summary, the thick line represents the region of the genome covered by the subclones. The marked restriction endonuclease sites represent the sites known to be contiguous with the subclones, determined from Southern blots of diagnostic digests of *C.elegans* genomic DNA probed with the $cpr{-}3$ cDNA (cm12b6). The extent of the region covered by the subclones which is known to be contiguous with the *C.elegans* genome is also indicated.

NOTES:

*EcoRV*: The *EcoRV* site is marked by an asterisk because its position was not determined from the original restriction mapping experiments. This site was originally identified from the DNA sequence data. Subsequently the presence and position of this single site within CL#034/1.7/KS+ was confirmed by double digests using *HindIII* and *EcoRV*. Later triple digests of CL#034/5/KS+ with *EcoRV*, *HindIII* and *SacII* demonstrated that no *EcoRV* site was present in this subclone.

The size of the subclones, the position of the restriction maps, the extent of the sequenced region and the sizes of the coding regions and introns are all drawn to the same scale (shown in the figure). The DNA sequence of the $cpr{-}3$ gene was aligned with the subclones, by aligning those restriction endonuclease sites identified by both restriction mapping and DNA sequencing.
A summary of the restriction map and DNA sequence data obtained from the subclones, ZK1055/6.8/KS- and ZK1055/2.8/KS+

The figure shows the position and size of the two subclones ZK1055/6.8/KS- and ZK1055/2.8/KS+ with respect to one another. The position of the restriction endonuclease sites determined by restriction mapping are also shown for each of these subclones. A summary of the DNA sequence of the cpr-4 gene, obtained from the two subclones, is shown above the subclones. This summary shows the orientation of the gene (5'→3', marked by the arrow), the coding regions of the gene (marked by black boxes), the introns (marked by the thin lines between boxes) and the mapped restriction endonuclease sites that were confirmed by sequencing. Above the DNA sequence summary, the thick line represents the region of the genome covered by the subclones. The marked restriction endonuclease sites represent the sites known to be contiguous with the subclones, determined from Southern blots of diagnostic digests of C.elegans genomic DNA probed with the cpr-4 cDNA (cm14e3). The extent of the region covered by the subclones which is known to be contiguous with the C.elegans genome is also indicated.

NOTES: The size of the subclones, the position of the restriction maps, the extent of the sequenced region and the sizes of the coding regions and introns are all drawn to the same scale, shown in the figure. The DNA sequence of the cpr-4 gene was aligned with the subclones, by aligning those restriction endonuclease sites identified by both restriction mapping and DNA sequencing.
Figure 3.27

A summary of the restriction map and DNA sequence data obtained from the subclones, W02B2/11/KS-, W02B2/3/KS+ and W02B2/1.6/KS+

The figure shows the position and size of the three subclones W02B2/11/KS-, W02B2/3/KS+ and W02B2/1.6/KS+ with respect to one another. The position of the restriction endonuclease sites determined by restriction mapping are also shown for each of these subclones. A summary of the DNA sequence of the cpr-5 gene, obtained from the two subclones, is shown above the subclones. This summary shows the orientation of the gene (5'-3', marked by the arrow), the coding regions of the gene (marked by black boxes), the introns (marked by the thin lines between boxes) and the mapped restriction endonuclease sites that were confirmed by sequencing. Above the DNA sequence summary, the thick line represents the region of the genome covered by the subclones. The marked restriction endonuclease sites represent the sites known to be contiguous with the subclones, determined from Southern blots of diagnostic digests of C.elegans genomic DNA probed with the cpr-5 cDNA (cm04d10). The extent of the region covered by the subclones which is known to be contiguous with the C.elegans genome is also indicated.

NOTES:

*BssHII: The BssHII site marked with the asterisk was originally identified from the DNA sequence data. The presence and position of a single BssHII site in W02B2/3/KS+ site was subsequently confirmed by BssHII digests of this subclone. It is not known whether BssHII sites are also present in the region covered by the cloned insert of W02B2/1.6/KS+ since this was not tested.

The size of the subclones, the position of the restriction maps, the extent of the sequenced region and the sizes of the coding regions and introns are all drawn to the same scale, shown in the figure. The DNA sequence of the cpr-5 gene was aligned with the subclones, by aligning those restriction endonuclease sites identified by both restriction mapping and DNA sequencing.
Figure 3.28

A summary of the restriction map and DNA sequence data obtained from the subclones, C25B8/10.7/KS-, C25B8/3.7/KS+ and C25B8/3/KS+

The figure shows the position and size of the three subclones C25B8/10.7/KS-, C25B8/3.7/KS+ and C25B8/3/KS+ with respect to one another. The position of the restriction endonuclease sites determined by restriction mapping are also shown for each of these clones. The striped boxes indicate the presence of approximately 700 bp of pJB8 cosmid vector sequence. A summary of the DNA sequence of the cpr-6 gene, obtained from the three subclones, is shown above the subclones. This summary shows the orientation of the gene (5'-3', marked by the arrow), the coding regions of the gene (marked by black boxes), the introns (marked by the thin lines between boxes) and the mapped restriction endonuclease sites that were confirmed by sequencing, with the exception of the HindIII sites which will be discussed below. Above the DNA sequence summary, the thick line represents the region of the genome covered by the subclones. The marked restriction endonuclease sites represent the sites known to be contiguous with the subclones, determined from Southern blots of diagnostic digests of C.elegans genomic DNA probed with the cpr-6 cDNA (cm01a5). The extent of the region covered by the subclones which is known to be contiguous with the C.elegans genome is also indicated.

NOTES:

**HindIII sites:** The HindIII sites marked by a single asterisk were originally identified from the DNA sequence of the cpr-6 gene. The position of the HindIII site marked by two asterisks in C25B8/3.7/KS+ was subsequently determined by restriction mapping using single (HindIII) and double (HindIII/SalI) digests. The HindIII site marked by three asterisks was determined from Southern blots of diagnostic digests of both C.elegans genomic DNA and C25B8/10.7/KS- DNA probed with the cpr-6 cDNA (cm01a5) which revealed there was a previously undetermined HindIII site 2.0 kb downstream of the most 3' HindIII site within the cpr-6 gene.

The size of the subclones, the position of the restriction maps, the extent of the sequenced region and the sizes of the coding regions and introns are all drawn to the same scale, shown in the figure. The DNA sequence of the cpr-6 gene was aligned with the subclones, by aligning those restriction endonuclease sites identified by both restriction mapping and DNA sequencing.
Figure 3.29

A Southern blot of CL#034 DNA digested with a variety of restriction endonucleases and electrophoresed through a 0.8% agarose gel

The Southern blot was hybridised with the gel purified 250 bp \textit{HindIII}/\textit{SacII} fragment from CL#034/5/\textit{KS+} (see restriction map, Figure 3.20) and washed at high stringency prior to exposure with medical X-ray film.

Lane 1: \textit{HindIII}
Lane 2: \textit{ClaI}
Lane 3: \textit{XbaI}
Lane 4: \textit{PstI}
Lane 5: \textit{EcoRI}
Figure 3.30

A Southern blot of C25B8/10.7/KS- DNA digested with *PstI* and *SacI* and electrophoresed through a 1.0% agarose gel.

The Southern blot was hybridised with pBluescript II KS- DNA and washed at high stringency prior to exposure with medical X-ray film.
The restriction mapping data used to determine the position of the *HindIII* site upstream of the *cpr-6* gene in the C25B8/3.7/KS+ subclone

A. A summary of the positions of the *HindIII* sites in the C25B8/3.7/KS+ subclone

The *HindIII* sites were determined by nucleotide sequencing of the *cpr-6* gene and by restriction mapping using *HindIII* and *HindIII/Sall* digests.

B. The restriction map data used to determine the *HindIII* site upstream of the *cpr-6* gene

Note: Within the sequenced region of C25B8/3.7/KS+, the boxes indicate coding regions of the *cpr-6* gene and the lines between these boxes indicate introns.
The position of this HindIII site was determined from HindIII and HindIII/SalI digests (shown in B). The scale in B is 0.1 Kbp.

Restriction Fragments: H = HindIII, S = SalI, PstI

Sequenced Region of C258B3.7/KS+ Polylinker + Pheledesmin II KS + HindIII Site in HindIII Site

Digests of C258B3.7/KS+ (shown in B)

PBR Vector Sequence (0.7 Kbp)
Diagnostic digests performed with *C. elegans* genomic DNA in order to determine whether the CL#034/5/KS+ and CL#034/1.7/KS+ subclones containing the *cpr-3* gene are contiguous with the *C. elegans* genome.

A. A Southern blot of EcoRV/SacII/HindIII digested CL#034/5/KS+ DNA electrophoresed through a 0.7% agarose gel. The Southern blot was hybridised with the *cpr-3* cDNA (cm12b6) and washed at high stringency prior to exposure with medical X-ray film.

B. A Southern blot of *C. elegans* Bristol N2 genomic DNA digested with a variety of restriction endonucleases and electrophoresed through a 0.7% agarose gel. The Southern blot was hybridised with the *cpr-3* cDNA (cm12b6) and washed at high stringency prior to exposure with medical X-ray film.

Lane 1:  
HindIII/XhoI

Lane 2:  
XhoI/SacII/EcoRV

Lane 3:  
XhoI/SacII/HindIII

Lane 4:  
XhoI/EcoRV/HindIII

C. Table summarising the expected sizes of the hybridising fragments generated from the diagnostic digests in B
Figure 3.33

Diagnostic digests performed with *C. elegans* genomic DNA in order to determine whether the ZK1055/6.8/KS- and ZK1055/2.8/KS+ subclones containing the *cpr-4* gene are contiguous with the *C. elegans* genome.

A. A Southern blot of *C. elegans* Bristol N2 genomic DNA digested with a variety of restriction endonucleases and electrophoresed through a 0.7% agarose gel. The Southern blot was hybridised with the *cpr-4* cDNA (cm14e3) and washed at high stringency prior to exposure with medical X-ray film.

**Lane 1:**  \( HindIII \)

**Lane 2:**  \( HindIII/BamHI \)

**Lane 3:**  \( PstI/BamHI \)

B. Table summarising the expected sizes of the hybridising fragments generated from the diagnostic digests in A
### A

![Image of gel electrophoresis](image-url)

### B

<table>
<thead>
<tr>
<th>Lane Number</th>
<th>Digest</th>
<th>Expected Fragment Sizes</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>HindIII</td>
<td>6.8 kb</td>
</tr>
<tr>
<td>2</td>
<td>HindIII / BamHII</td>
<td>2.9 kb, 4.1 kb</td>
</tr>
<tr>
<td>3</td>
<td>PstI / BamHII</td>
<td>1.4 kb, 2.0 kb</td>
</tr>
</tbody>
</table>
Diagnostic digests performed with *C. elegans* genomic DNA in order to determine whether the W02B2/3/KS+ and W02B2/1.6/KS+ subclones containing the cpr-5 gene are contiguous with the *C. elegans* genome.

A. A Southern blot of *C. elegans* Bristol N2 genomic DNA digested with a variety of restriction endonucleases and electrophoresed through a 0.7% agarose gel. The Southern blot was hybridised with the cpr-5 cDNA (cm04d10) and washed at high stringency prior to exposure with medical X-ray film.

Lane 1:    *BamHI/PstI*
Lane 2:    *BamHI*
Lane 3:    *BamHI/PstI/XbaI*

B. Table summarising the expected sizes of the hybridising fragments generated from the diagnostic digests in A
## B

<table>
<thead>
<tr>
<th>Lane Number</th>
<th>Digest</th>
<th>Expected Fragment Sizes</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>BamHI / PstI</td>
<td>3.05 kb, 1.6 kb kb</td>
</tr>
<tr>
<td>2</td>
<td>BamHI</td>
<td>3.05 kb, and a fragment larger than 1.6 kb</td>
</tr>
<tr>
<td>3</td>
<td>BamHI / PstI / Xbal</td>
<td>1.6 kb, 1.4 kb</td>
</tr>
</tbody>
</table>
Diagnostic digests performed with *C. elegans* genomic DNA in order to determine the extent of the region of the C25B8/10.7/KS- subclone containing the *cpr-6* gene which is contiguous with the *C. elegans* genome.

**A.** A Southern blot of C25B8/10.7/KS- DNA digested with a variety of restriction endonucleases and electrophoresed through a 0.7% agarose gel. The Southern blot was hybridised with the *cpr-6* cDNA (cm01a5) and washed at high stringency prior to exposure with medical X-ray film.

**Lane 1:** *HindIII*
**Lane 2:** *HindIII/BamHI*
**Lane 3:** *HindIII/SalI*

**B.** A Southern blot of *C. elegans* Bristol N2 genomic DNA digested with a variety of restriction endonucleases and electrophoresed through a 0.7% agarose gel. The Southern blot was hybridised with the *cpr-6* cDNA (cm01a5) and washed at high stringency prior to exposure with medical X-ray film.

**Lane 1:** *HindIII*
**Lane 2:** *HindIII/BamHI*
**Lane 3:** *HindIII/SalI*

**C.** Table summarising the expected sizes of the hybridising fragments obtained from the diagnostic digests in A and B
### C

<table>
<thead>
<tr>
<th>Lane Number</th>
<th>Digest</th>
<th>Expected Fragment Sizes</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>HindIII</td>
<td>0.5 kb, 0.7 kb, 1.5 kb, fragment of unknown size</td>
</tr>
<tr>
<td>2</td>
<td>HindIII / BamHI</td>
<td>0.5 kb, 0.63 kb, 1.5 kb, fragment of unknown size</td>
</tr>
<tr>
<td>3</td>
<td>HindIII / SalI</td>
<td>0.5 kb, 0.7 kb, 0.15 kb, fragment of unknown size</td>
</tr>
</tbody>
</table>
Chapter 4
Chapter 4
Characterisation of cpr-3, cpr-4, cpr-5 and cpr-6

4.1. Introduction

The isolation of four cathepsin B-like genes from *C. elegans* reported in Chapter 3, in addition to the previously isolated *cpr-1* gene (Ray and McKerrow, 1992), indicates that this free living nematode species possesses a cathepsin B-like multigene family comprising at least five members. Multigene families with the potential to encode cathepsin B-like enzymes have also been isolated from parasitic nematode and trematode species (Chapter 1, Section 1.9) but not from any other metazoan.

The presence of a cathepsin B-like multigene family in both parasitic and free living nematode species raises interesting questions as to the biological function, and degree of interaction, of the members of this gene family. Such a multigene family may have arisen as a result of a requirement for a large amount of cathepsin B-enzyme over a very short period of time. However, this seems unlikely since one might expect such a requirement to be fulfilled by a single cathepsin B-like gene under the control of a strong promoter. Therefore, a more likely explanation for the amplification of this multigene family in nematode species, may be the requirement for a range of cathepsin B-like enzymes with differing substrate specificities. The multicellular nature of nematode species may be an important factor in this respect, since enzymes with different substrate specificities may be required in different tissues, or in different compartments within tissues. Alternatively, such enzymes may be required at different stages during nematode development.

The subclones containing *cpr-3, cpr-4, cpr-5* and *cpr-6* (Chapter 3) were isolated using four partially sequenced cDNA clones. The partial sequence data available for the four cDNA clones were originally generated by only a single sequencing reaction on one strand of each cDNA clone. Consequently, the amount (approximately 300 bp) and reliability of this sequence data were limited. Therefore, I decided to sequence the cDNA and genomic clones completely. This approach covered several objectives: to confirm that the subclones isolated using the four cDNA clones (Chapter 3) contained the genes corresponding to these cDNA clones; to determine the predicted primary protein
structure of the four enzymes; to determine the genomic architecture of the four genes; to map the region of transcription initiation and to identify putative regulatory sequence elements in the 5' flanks of the four genes.

4.2. Results

4.2.1. Confirming the identity of cpr-3, cpr-4, cpr-5 and cpr-6

The cDNA clones were sequenced completely on both strands. The regions of each gene within the appropriate subclone that were also contained within the cDNA clones were sequenced on one strand. A comparison of the cDNA and coding genomic sequences revealed that the cDNA clones cm12b6, cm14e3, cm04d10 and cm01a5 shared complete base identity to their genes cpr-3, cpr-4, cpr-5 and cpr-6, present in the genomic clones CL#034, ZK1055, W02B2 and C25B8, respectively. These data formally indicated that the correct genes had been isolated using these cDNA clones as probes. Therefore, the cDNA clones cm12b6, cm14e3, cm04d10 and cm01a5 are subsequently referred to by their gene names, cpr-3, cpr-4, cpr-5 and cpr-6. A more detailed discussion of the data revealed by this comparison will be given later.

4.2.2. Sequencing the cDNA clones of cpr-3, cpr-4, cpr-5 and cpr-6

Each of the four cDNA clones were sequenced on both strands. The results of this analysis indicated that all four clones possess intact 3' ends with poly(A) tails of at least eleven residues. The cDNA clones of cpr-3 and cpr-4 both possessed canonical cleavage/poly(A) signals (AAUAAA) 14 bp and 15 bp upstream of the poly(A) tail. The two other cDNAs of cpr-5 and cpr-6 possess variations to this signal of GAUAAA and AAUGAA respectively, 15 bp and 13 bp upstream of the poly(A) tail, which probably serve as the cleavage/poly(A) signal. This conclusion is drawn from a survey of the 3' ends of 1303 C.elegans cDNA clones (T.Blumenthal, pers. comm.). This survey revealed that a significant number of the cDNA clones analysed have variations of the canonical cleavage/poly(A) signal upstream of the poly(A) tail. Two of the more common variations found are GAUAAA (4.8%) and AAGUAA (8.3%). The survey also
revealed that putative cleavage/poly(A) signals are most frequently found 11-15 bp upstream of the poly(A) tail. These observations lead T.Blumenthal (pers. comm.) to suggest that any of the sequence variations found with a frequency of over 1% in the correct locale are likely to be acting as cleavage/poly(A) signals.

The predicted amino acid sequence (discussed in detail in Section 4.2.7) indicated that the cDNA clones of cpr-3, cpr-4 and cpr-6 probably contained the ATG translation initiation codons, however the 5′ untranslated regions (UTRs) of these three clones were short (21 to 39bp), suggesting that these clones may be truncated at the 5′ end. The cDNA clone of cpr-5 is definitely truncated at the 5′ end, with no putative ATG initiation codon present.

4.2.3. Obtaining the full-length mature transcripts of cpr-3, cpr-4, cpr-5 and cpr-6

I decided to use the 5′ Rapid Amplification of cDNA Ends (5′ RACE) method to obtain the 5′ ends of the mature transcripts for each of the four genes. A summary of this protocol is shown in Figure 4.1, and the nucleotide sequences of all the primers used are given in Chapter 2, Section 2.11.3.2. Two pools of first strand cDNA were generated from C.elegans Bristol N2 mixed stage total RNA using pooled gene specific primers for each of the four genes cpr-3, cpr-4, cpr-5 and cpr-6 and Superscript II Reverse Transcriptase (GibcoBRL). The two pools of first strand cDNA were tailed using dCTP Terminal Deoxynucleotidyl Transferase (TdT) and amplified by 36 cycles of PCR using pooled nested gene specific PCR primers for each of the four genes and the GibcoBRL anchor primer, according to the GibcoBRL 5′ RACE protocol. Each of the two sets of amplification products generated were re-amplified in four independent PCR reactions, using the anchor primer, one nested gene specific PCR primer and 36 cycles of PCR. The results obtained using 5′ RACE for each of the four genes are show in Figure 4.2. The results indicated that the 5′ RACE reactions from the two pools of cDNA reproducibly generate the same size of predominant products for each gene which migrate as discrete bands when electrophoresed through a 1.5% agarose gel. The 5′ RACE experiments generated a single predominant size of product for cpr-3, cpr-4 and cpr-5, suggesting that the 5′ ends of the mature transcripts of these genes are predominantly of a single species. For cpr-6, two predominant sizes of 5′ RACE
product are generated, migrating approximately 60 bp apart, initially suggesting that there may be two predominant species of mature transcript derived from this gene. The predominant products were gel purified and cloned into pBluescript II KS+ by virtue of restriction enzyme sites present in the primers used for the final amplification reaction. A total of at least five 5' RACE clones, generated from the two cDNA populations, were sequenced on both strands for each gene. The nucleotide sequence of each 5' RACE clone was determined for the region outwith the 5' end of the appropriate cDNA clone, in addition to a 50 bp overlap to confirm identity with the cDNA clone sequence. The cloning sites at the 5' and 3' ends of each 5' RACE clone were then identified by scanning the remaining sequence of each clone. This allowed the size of each 5' RACE clone to be determined, and these sizes were compared to the expected sizes obtained from initial electrophoresis of the 5' RACE products (Figure 4.2). This comparison (Table 4.1) reveals that the sizes of the cloned 5' RACE products correlate well with the expected sizes determined from agarose gel electrophoresis of the 5' RACE products prior to cloning.

For cpr-6, two similarly abundant 5' RACE products were obtained, migrating approximately 60 bp apart. These products were gel purified together and cloned into pBluescript II KS+. Six of the cloned products were selected at random and sequenced. These results indicated that all six clones were of a similar size to each other, and corresponded to the smaller of the two products generated by 5' RACE (Table 4.1). This was a surprising result since the two 5' RACE products were of a similar abundance, and therefore one would expect both products to be represented in the 5' RACE clones selected for sequence analysis. These data suggested that the 5' end of the mature transcript of cpr-6 was also predominantly of a single species, and therefore the larger product generated after the second round of PCR was probably an artefact. The second round of PCR was repeated for cpr-6 using one of the two pools of amplified cDNA previously obtained from the first round PCR. The results indicated that only a single predominant 5' RACE product is generated (Figure 4.3), which is a similar size to the smaller of the two products previously observed for cpr-6 (Figure 4.2). These data support the assumption that the larger 5' RACE product previously observed for cpr-6 was indeed an artefact.
The results obtained from DNA sequencing the cloned 5' RACE products of cpr-3, cpr-4, cpr-5 and cpr-6 are shown in Figures 4.4, 4.5, 4.6 and 4.7, respectively. Some of the products showed base substitutions at various points in the sequence and these have been indicated. These substitutions were assumed to have occurred during the PCR reaction. The Taq DNA polymerase used in the PCR reaction does not have any proof reading ability, and therefore base mismatches may occur during the amplification step. Furthermore, the frequency of such errors increases with the number of PCR cycles used. In this instance, the use of a large number of PCR cycles (72 in total) was therefore likely to generate several errors. In order to overcome this a consensus sequence was generated from all the sequence data obtained from the cloned 5' RACE products, for each gene. A comparison of the consensus sequence obtained from the 5' RACE products for each of the genes cpr-3, cpr-4, cpr-5 and cpr-6 to the appropriate cDNA revealed complete base identity in the region of overlap, with the exception of the 5' most base of each of the cDNA clones. This base was immediately downstream of the SacI site used to clone these cDNAs during initial preparation of the library (Palazzolo et al., 1990) and was therefore probably an artefact. The library was originally prepared by size fractionation of poly(A)+ RNA, generating first strand cDNA with an oligo(dT)-ApaI primer-adaptor and 5-methyl-dCTP instead of dCTP. Second strand cDNA was then generated using unmethylated dCTP, to generate hemimethylated cDNA. A SacI linker was then ligated to the 5' end of the cDNA and the cDNAs digested with ApaI and SacI (any endogenous sites being protected by the methylation) and ligated into the lambda SHLX2 phage vector. All the cDNA clones possess a guanine residue immediately downstream of the SacI site, and it is this residue in addition to the SacI site, that mismatches with the sequence data obtained from the 5' RACE clones. This residue was therefore probably added to each of the cDNA clones during the generation of the cDNA library, as part of the SacI linker.

The sequence data obtained from the cloned 5' RACE products of cpr-3 (Figure 4.4), cpr-4 (Figure 4.5), cpr-5 (Figure 4.6), and cpr-6 (Figure 4.7) all generated additional sequences outwith the 5' end of the corresponding cDNA clone. For cpr-5 (Figure 4.6), the 5' RACE products generated a single ATG codon which was in frame with the predicted amino acid sequence of the corresponding cDNA clone. This ATG codon probably represents the translation initiation codon, since eukaryotic translation is
thought to occur at the first ATG codon of a given transcript (Kozak, 1983). For 

\textit{cpr-4} (Figure 4.5), and \textit{cpr-6} (Figure 4.7), translation is most likely to be initiated from the 

ATG codons previously identified in the cDNA clones (Section 4.2.2), since the 

sequenced 5′ RACE clones from these two genes did not reveal any additional ATG 

codons in the 5′ UTRs of their mature transcripts. For \textit{cpr-3} (Figure 4.4), the ATG 

codon originally identified by sequencing the cDNA clone of this gene (4.2.2) is also the 

putative translation initiation codon. However, one of the eight sequenced 5′ RACE 

clones of \textit{cpr-3} possesses an additional ATG codon upstream of, and in-frame with, this 

putative translation initiation codon. This upstream ATG codon is unlikely to be the 

translation initiator since an in-frame translation termination codon separates it from the 

putative translation initiator. Thus, translation initiation from this ATG codon would 

only generate a 22 amino acid peptide. It is possible that this peptide may have a 

regulatory role such as an enzyme inhibitor, or by preventing synthesis of the enzyme 

from this transcript. However, the physiological significance of this additional ATG 

codon must be questioned since only one of the eight sequenced 5′ RACE clones of \textit{cpr-3} possesses it, suggesting that this ATG codon is not present in the majority of mature 

transcripts generated from the \textit{cpr-3} gene.

The predicted amino acid sequences, generated from the predicted translation 

initiation codons for each of the four genes, all generate very similar primary protein 

structures (Figure 4.8, discussed later). In particular, all four possess highly hydrophobic 

putative pre-enzyme domains which have been identified in all other cathepsin B-like 

genes sequenced to date. The generation of four structurally similar primary protein 

sequences from the mature transcripts of the four genes, \textit{cpr-3}, \textit{cpr-4}, \textit{cpr-5} and \textit{cpr-6}, 

using the chosen ATG codons, strongly supports their role as translation initiators.

The cloned 5′ RACE products for all four genes, \textit{cpr-3}, \textit{cpr-4}, \textit{cpr-5} and \textit{cpr-6}, 

all indicated short 5′ UTRs. A comparison of the consensus sequence of the 5′ UTRs, 

up to and including the ATG translation initiation codon, to the 5′ flank sequence 

obtained from the genomic clones for each of the four genes revealed complete base 

identity along their entire length (Figure 4.9). This indicates that none of the 

predominant transcripts from \textit{cpr-3}, \textit{cpr-4}, \textit{cpr-5} and \textit{cpr-6} are trans-spliced to either of 

the two 22 nucleotide splice leaders of \textit{C.elegans}, SL1 and SL2 (Appendix A). This is 

unusual since recent reports suggest that approximately 70% of \textit{C.elegans} genes are
trans-spliced to one of these two splice leaders (Zorio et al., 1994). However, these data are in agreement with data obtained from the previously isolated cathepsin B-like gene of *C. elegans*, *cpr-1*, which also suggest that this gene is not trans-spliced (Ray and McKerrow, 1992). These data suggest that the members comprising the cathepsin B-like multigene family of *C. elegans* all lack this aspect of transcript maturation. In *C. elegans*, some genes have been shown to be transcribed polycistronically in clusters resembling bacterial operons (Spieth et al., 1993). Mature, monocistronic messages are subsequently generated from the polycistronic message by trans-splicing events. A recent study of potential operons revealed by the *C. elegans* genome sequencing project suggested that approximately 26% of genes may be organised into operons (Zorio et al., 1994). The authors demonstrated that, of the operons studied, the first gene was either trans-spliced to SL1 or not trans-spliced at all while the downstream genes were exclusively trans-spliced to SL2, or a mixture of SL1 and SL2. These data suggest that if any of the *cpr-1, cpr-3, cpr-4, cpr-5* or *cpr-6* genes are transcribed from *C. elegans* operons, they will represent the first gene in their operon. The two genes, *cpr-1* and *cpr-3*, are closely linked on chromosome V (Chapter 3, Section 3.2.13) suggesting that, of all the cathepsin B-like genes identified to date, these two are the most likely to be co-transcribed from the same operon. However, the data suggest that neither of these two genes are trans-spliced and are therefore unlikely to represent the downstream gene of a *C. elegans* operon. This suggests that these two genes cannot be transcribed from the same operon, though they could each represent the first gene of two different operons.

4.2.4. Analysis of the full-length mature transcripts of *cpr-3, cpr-4, cpr-5 cpr-6*

The full length mature transcripts were obtained from the combined sequence data of the cDNA and 5’ RACE clones for each gene. The sizes of the transcripts were calculated from the largest 5’ UTR generated by more than one 5’ RACE clone (for *cpr-3, cpr-4* and *cpr-6*). For *cpr-5*, the largest 5’ UTR was used because none of the sequenced 5’ RACE clones terminated at the same position for this gene. The sizes were found to be, 1,329 bases, 1,120 bases, 1,152 bases and 1,260 bases for *cpr-3, cpr-4, cpr-5* and *cpr-6* respectively. The open reading frames of *cpr-3, cpr-4, cpr-5* and *cpr-6* encode proteins of 370, 335, 344 and 379 amino acids respectively. Subsequent analysis
of these amino acid sequences indicated that these four genes encode cathepsin B-like cysteine protease enzymes that are similar to, but distinct from, each other and the 329 amino acid cathepsin B-like gene product of the previously isolated \textit{cpr-1} gene, a more detailed comparison of these sequences will be given later. The mature transcripts of \textit{cpr-1}, \textit{cpr-3}, \textit{cpr-4}, \textit{cpr-5} and \textit{cpr-6} all shared similar features. As discussed in the previous section, none of the transcripts are trans-spliced. Furthermore, the 5' untranslated sequences of these five genes are all relatively small. The 5' RACE experiments performed for \textit{cpr-3}, \textit{cpr-4}, \textit{cpr-5} and \textit{cpr-6} generated several 5' UTRs of slightly different length for each gene (Figures 4.4, 4.5, 4.6 and 4.7). The size ranges for the 5' UTRs of \textit{cpr-3}, \textit{cpr-4}, \textit{cpr-5} and \textit{cpr-6} obtained from the 5' RACE experiments are 39-92 bases, 33-39 bases, 14-29 bases and 29-46 bases, respectively. Only the mature transcript of \textit{cpr-3} possesses a 5' UTR substantially larger than is found for the other three genes and \textit{cpr-1}. It should be noted that the wide range of 5' UTR sizes found for \textit{cpr-3} are slightly misleading since the extremities of these sizes (92 bases and 39 bases) were each generated by single 5' RACE clones. The majority of 5' RACE clones for \textit{cpr-3} generated 5' UTRs of 60 bases (3 clones) and 55 bases (3 clones) which were only slightly larger than the 5' UTRs found for the other \textit{C. elegans} cathepsin B-like genes. The 5' UTRs of the mature transcripts of \textit{cpr-3}, \textit{cpr-4}, \textit{cpr-5} and \textit{cpr-6} are a similar size to the 28 bp 5' UTR of \textit{cpr-1}, determined using primer extension analysis (Ray and McKerrow, 1992). These data suggest that short 5' UTRs may be a common feature of the mature transcripts of all \textit{C. elegans} cathepsin B-like genes.

The FOLDRNA program of the GCG package was used to predict RNA secondary structures for the 5' and 3' UTRs of \textit{cpr-3}, \textit{cpr-4}, \textit{cpr-5}, \textit{cpr-6} and \textit{cpr-1} in order to determine whether there were any stable stem-loop structures in the 5' UTRs that could potentially perform regulatory roles for these genes. Only the 3' UTR of \textit{cpr-3} generated a stem-loop structure that showed significant stability (-20.1 kilocalories at 37°C, Figure 4.10). This is not surprising since the 3' UTR of \textit{cpr-3} (156 bases) is notably larger than the 3' UTRs of \textit{cpr-4} (76 bases), \textit{cpr-5} (88 bases), \textit{cpr-6} (75 bases) and \textit{cpr-1} (62 bases). The stable stem-loop structure generated by the 3' UTR of the \textit{cpr-3} transcript may represent a mechanism by which \textit{cpr-3} expression is post-transcriptionally regulated (discussed in Section 4.3.4.1).
4.2.5. The gene architectures of cpr-3, cpr-4, cpr-5 and cpr-6

The genomic copies of cpr-3, cpr-4, cpr-5 and cpr-6 were sequenced using custom made primers (Chapter 2, Section 2.11.3.2) and template DNA from the following subclones: CL#034/1.7/KS+, CL#034/3/KS+ and CL#034/5/KS+ for cpr-3; ZK1055/2.8/KS+ and ZK1055/6.8/KS- for cpr-4; W02B2/1.6/KS+ and W02B2/3/KS+ for cpr-5; C25B8/3/KS+, C25B8/3.7/KS+ and C25B8/10.7/KS- for cpr-6. A summary of the position of these subclones with respect to one another and with respect to the appropriate gene is shown in Chapter 3, Figures 3.25, 3.26, 3.27 and 3.28. It should be noted that W02B2/1.6/KS+ and W02B2/3/KS+ are adjacent, not overlapping subclones. These clones were shown to be adjacent by comparison of the nucleotide sequence of the cpr-5 gene, which spans these two clones, to the nucleotide sequence of the cDNA clone of the cpr-5 gene (cm04d110).

The coding sequences of the genomic clones of cpr-3, cpr-4, cpr-5 and cpr-6 were determined by sequencing one strand and confirming identity with the corresponding cDNA clone. Those regions of the genomic sequence not contained within the cDNA clones (introns and 5' and 3' flanks), were sequenced on both strands. A comparison of the nucleotide sequence of the mature transcripts (generated from the combined cDNA clone and 5' RACE product sequence data) of each of the four genes to the nucleotide sequence of their genomic clones, reveals complete base identity interrupted by introns. The annotated DNA sequences of cpr-3, cpr-4, cpr-5 and cpr-6 generated by these comparisons are shown in Figures 4.11, 4.12, 4.13 and 4.14, respectively. All the introns revealed within cpr-3, cpr-4, cpr-5 and cpr-6 possess the conserved splice donor (GT) and splice acceptor (AG) sites expected for C.elegans introns (Emmons, 1988). However, the size, number and position of these introns vary considerably between the genes giving rise to quite different gene architectures, which are shown schematically in Figure 4.15. Intrion size ranges from 43 bp (cpr-6) to 406 bp (cpr-3) while number varies from one (cpr-4) to seven (cpr-6). The range in size and number of introns for the genes cpr-3, cpr-4, cpr-5 and cpr-6, are well within the limits expected for C.elegans. Early analysis of 54 different C.elegans introns revealed that three-quarters were between 45 and 59 bases long (Emmons, 1988). Experiments performed with the rabbit B-globin gene suggest introns shorter than 80 bp may not be
effectively spliced in higher eukaryotes (Wieringa et al., 1984). This suggests that the mechanism employed for splicing C.elegans introns is different from that for splicing introns from higher eukaryotes.

The position of the introns were also compared at the amino acid sequence level (Figure 4.8). In most cases intron position at the amino acid level is not conserved, however there are two exceptions. The cpr-4 and cpr-6 genes share a common intron site immediately after the ATG start codon (position 1) with the previously characterised gene cpr-1. The cpr-3 and cpr-5 genes share a common intron site with cpr-1 at position 261 (Figure 4.8). These intron positions are not only conserved with respect to the amino acid sequence but also with respect to the exact nucleotides at which intron / exon boundaries occur within the codons. The identification of two conserved sites of intron position with respect to the amino acid and nucleotide sequences of cpr-1, cpr-3, cpr-4, cpr-5 and cpr-6, suggest that these genes share a common ancestor. However, the different gene architectures, in terms of the position, number and size of the introns, suggest that these genes are highly diverged.

The COMPARE and DOTPLOT programs of the GCG package were used to identify homologous sequences from different regions of the same gene which might represent repeated sequences, by comparing cpr-1, cpr-3, cpr-4, cpr-5 and cpr-6 to themselves. These programs were also used to identify regions of homology between the genes by comparing the nucleotide sequences of cpr-1, cpr-3, cpr-4, cpr-5 and cpr-6 to each other. In the latter case, the COMPARE and DOTPLOT programs were necessary because the DNA sequences of each of the five genes are sufficiently diverged such that the GAP program of the GCG package (which aligns sequences by introducing gaps) was unable to generate any alignments between the different genes.

The COMPARE program works by scanning the sequence of one gene for homology to a window of sequence taken from the same, or another gene. The window of sequence is then moved along that gene in a stepwise manner and the process repeated. Regions of homology are then scored on a graph as a dot, using the DOTPLOT program. Successive regions of homology therefore give rise to successive dots which generate lines whose length are proportional to the length of the stretch of homology.
DOTPLOT comparisons of each of the genes *cpr-1, cpr-3, cpr-4, cpr-5* and *cpr-6* to themselves did not reveal any significant homologies between different regions of the same gene, except for *cpr-5*. For this gene, extensive homology was found in the first 80 bp of the 5′ flank of the sequenced gene. A study of this region revealed several repetitions of two hexamer sequences, TTATGC and TTAGGC (Figure 4.16). Seven repeats of TTATGC and four of TTAGGC were identified. However, further repeats containing one base pair mismatches to these sequences were also identified interspersed amongst these perfect repeats. The TTAGGC repeat identified is thought to be the telomeric hexamer of *C. elegans*. Recently the complete set of all 12 telomeres of *C. elegans* has been analysed and each telomere shown to comprise 4-6 kb of this repeated hexamer (F.Muller, pers. comm.). Interstitial repeats of TTAGGC have also been found in *C. elegans*, and these tend to be clustered in the terminal 30% of chromosomes (Cangiano and La Volpe, 1993). In such cases, degenerated copies of the telomeric hexamer have been observed, interspersed among perfect copies of this repeat sequence (Cangiano and La Volpe, 1993). The type and pattern of repeat sequence found in the 5′ flank of *cpr-5* suggests that this gene is closely associated with a region of interstitial telomeric repeats. This is not surprising since *cpr-5* has been mapped to the left end of chromosome V (Chapter 3, Section 3.2.13), where many interstitial copies of the telomeric repeat appear to be clustered. It should be noted that this region of the 5′ flank of *cpr-5* was not sequenced on both strands because it was not possible to generate a specific primer, required for sequencing the second strand within the region of repetitive sequence. Therefore, it was not possible to determine whether the degenerated copies of the two types of hexamer repeat found were genuine or a result of sequencing inaccuracies. Though the two types of hexamer repeat only differ themselves by one nucleotide residue, this variation is unlikely to be a result of sequencing inaccuracies because several copies of each type of repeat were observed.

The COMPARE and DOTPLOT programs identified several regions of homology between *cpr-1, cpr-3, cpr-4, cpr-5* and *cpr-6*. A summary of the data obtained from these dotplots (the position and length of the homologies) is given in Table 4.2. When these positions are compared to the annotated sequence data of *cpr-3, cpr-4, cpr-5* and *cpr-6* (Figures 4.11, 4.12, 4.13 and 4.14, respectively) it becomes apparent that these homologies are within the regions of the genes coding for the
catalytic residues (cysteine, histidine and asparagine), and the amino acid residues surrounding these sites. These data suggested that these genes are sufficiently diverged such that no significant DNA homology is present between introns, or exons outwith the regions which code for the catalytic domains of their gene products. The stretches of homology between genes, identified by the dotplots and summarised in Table 4.2, were compared to determine the degree of homology to each other using the GAP program of the GCG package. Though the GAP program was unable to generate alignments between the complete sequences of cpr-1, cpr-3, cpr-4, cpr-5 and cpr-6, it was able to generate alignments between the regions of homology identified from the dotplots. Presumably, this was a result of the fact that much shorter DNA sequences were being compared and that these sequences had already been selected on the basis of increased homology to one another using the COMPARE and DOTPLOT programs. The results of the GAP program analysis (summarised in Table 4.3) demonstrate that whilst all the genes share regions of significant homology to each other (80.8% - 100%), the length of these homologies are very small, the maximum being 29 bp of continuous homology between cpr-1 and cpr-4. The lack of any continuous homology over 29 bp between any two genes suggested that the genes cpr-1, cpr-3, cpr-4, cpr-5 and cpr-6 are highly diverged from each other at the nucleotide level. Indeed, the degree of divergence is sufficiently high to suggest that detection of one gene, using another as a hybridisation probe, would be difficult, or impossible.

Restriction endonuclease fingerprinting previously demonstrated that the clones containing cpr-1 and cpr-3 are very closely linked, with the two clones mapping on top of one another (Chapter 3, Section 3.2.13). Therefore, the COMPARE and DOTPLOT programs were used to compare the 5' and 3' flanks of cpr-3 to those of cpr-1 (in both orientations). This comparison did not identify any regions of homology between the 5' and 3' flanks of cpr-1 and cpr-3, indicating that neither of these two genes possess sequences of the other in the sequenced regions of their 5' and 3' flanks.

4.2.6. Putative regulatory elements in the 5' flanks of cpr-3, cpr-4, cpr-5 and cpr-6

5' RACE was used initially to obtain the predominant full size transcripts of the genes cpr-3, cpr-4, cpr-5 and cpr-6. However, the discovery that these genes were not
trans-spliced to either SL1 or SL2, allowed the 5' RACE results to be used to map the general region of transcription initiation for each of the genes. The 5' UTRs defined by the cloned 5' RACE products, up to and including the ATG translation initiation codon, were compared to approximately 300 bp of 5' flank sequence of the appropriate gene. This comparison demonstrated that the 5' ends of the RACE clones from each gene clustered to a small region of the 5' flank of the appropriate gene (Figure 4.17). This tight clustering suggests that transcription initiation occurs within one region of the 5' flank of each of the four genes. The only slight exception to this is in the case of *cpr-3*, where the 5' ends of two 5' RACE products were separated by 52 bases. In this example, however, the 5' ends of the six other 5' RACE clones for *cpr-3* mapped to two positions five bases apart. These data suggested that transcription initiation occurs predominantly from a single region, but that spurious transcription initiation may also occur outwith this area.

A survey of the 5' flank sequences of the four genes upstream of the regions to which the 5' ends of the 5' RACE clones mapped, revealed clustering of several putative regulatory elements. In two cases, *cpr-4* and *cpr-6*, putative TATA box (TATAWAW) promoter elements were identified, 25 and 26 bases respectively, upstream of the point to which most of their 5' RACE clones mapped (Figure 4.17). Since TATA boxes are generally found 25 bp upstream of the region of transcription initiation, these results suggest that the TATA sequences identified may represent functional promoter elements. For *cpr-3* and *cpr-5*, no such potential TATA box promoter sequences were identified indicating that a TATA-less promoter must be involved in transcription of these two genes. All four genes possess sequences with homology to the motif recognised by the GATA family of transcription factors. The 5' flank sequences of *cpr-3*, *cpr-4*, *cpr-5* and *cpr-6* contain four, five, two and two copies, respectively, of the classic (WGATAR, Figure 4.17) sequence recognised by the GATA family of transcription factors, plus additional related variants which might also bind this family of transcription factors (Ko and Engel, 1993; Merika and Orkin, 1993).
4.2.7. *cpr-3, cpr-4, cpr-5* and *cpr-6* encode proteins with homology to cathepsin B

The predicted amino acid sequences of *cpr-3, cpr-4, cpr-5* and *cpr-6*, and that of *cpr-1*, most closely resemble the cathepsin B cysteine proteases, as typified by human cathepsin B (Chan *et al.*, 1986). The predicted amino acid sequences of these six genes are compared in Figure 4.8. Each of the six genes possesses a short (16 - 19 residue) N-terminal hydrophobic domain. This hydrophobic domain is the candidate signal prepeptide region which may be responsible for directing the preproenzymes to the lumen of the endoplasmic reticulum during translation by the ribosomes (Docherty and Steiner, 1982). This domain is followed by a diverged region of 62 - 87 amino acids which is thought to represent the proregion of the enzymes. Such divergence in the proregion has also been noted between human cathepsin B and its homologues in the mouse and the rat (Chan *et al.*, 1986) and between the CP-1 and AC-2 cathepsin B-like enzymes of *H. contortus* and *O. ostertagi* (Pratt *et al.*, 1992b). The function of the proregion is unclear, however removal of this region is thought to be necessary for activity of the mature enzyme. Extracellular, stable, high molecular mass forms of recombinant human cathepsin B have been isolated from yeast. These forms appear to represent non-covalent complexes between the mature cathepsin B enzyme and its propeptide (Mach *et al.*, 1994b). The authors also reported that this complex does not exhibit a significant ability to cleave protein substrates. Furthermore, experiments using a synthetic rat cathepsin B propeptide have demonstrated that this propeptide is a potent reversible inhibitor of the rat cathepsin B enzyme (Fox *et al.*, 1992). The proregion of all six of the predicted protein products compared lack the interspersed ERFNIN motif. This motif has been identified in the proregion of a number of cysteine proteases but is absent from the proregion of all cathepsin B-like enzymes analysed (Karrer *et al.*, 1993). The predicted region of cleavage of the propeptide from the mature enzyme of *cpr-3, cpr-4, cpr-5* and *cpr-6* is indicated in Figure 4.8. This position is based on the predicted site of cleavage for *cpr-1* (Ray and McKerrow, 1992) and the known site of cleavage for human cathepsin B (Chan *et al.*, 1986) which was determined by comparison of the predicted amino acid sequence obtained from a human cathepsin B cDNA and amino acid sequencing of mature human cathepsin B. The predicted site of cleavage of the propeptide from the mature enzyme of *cpr-3, cpr-4, cpr-5* and *cpr-6* is very near to the
site of cleavage for AC-4 of H. contortus, determined by N-terminal protein sequencing (Pratt et al., 1992a).

The predicted mature forms of the enzymes show much higher levels of homology than the prepropeptides, as can be seen from the number of amino acid residues conserved between the prepro- regions and the predicted mature regions of the six enzymes (Figure 4.8). The percentage similarities and identities between the prepro and predicted mature forms of the enzyme products of cpr-1, cpr-3, cpr-4, cpr-5, cpr-6 and human cathepsin B are shown in Table 4.4. These data also demonstrate that the predicted mature forms of the enzymes are more homologous to one another (51.0% - 77.3% identity) than the preproenzyme forms (43.1% - 67.6% identity). However, the data summarised in Table 4.4 also demonstrate that the putative enzymes encoded by cpr-1, cpr-3, cpr-4, cpr-5 and cpr-6 are not highly homologous, since (with the exception of cpr-4 and cpr-5) these putative enzymes do not show significantly more homology to one another than to human cathepsin B.

Despite the diverged nature of the cathepsin B-like enzymes encoded by cpr-1, cpr-3, cpr-4, cpr-5 and cpr-6, the predicted mature forms of these enzymes possess the characteristics expected of cathepsin B enzymes. The three definitive catalytic residues, Cysteine (C), Histidine (H) and Asparagine (N), are highly conserved, as are the amino acid residues surrounding these sites. In addition, the position of these three catalytic residues, with the cysteine residue situated towards the N-terminus and the histidine and asparagine residues situated towards the C-terminus, are highly conserved between the six enzymes. Fourteen of the cysteine residues in the predicted mature enzymes can be aligned with few gaps. Since cysteine residues are required for the production of disulphide bonds, necessary for generating the tertiary structure of the proteins, the ability to align such a high number of these residues with a minimal number of gaps suggests these six gene products all have similar tertiary structures. These data indicate that the four genes isolated, cpr-3, cpr-4, cpr-5 and cpr-6, encode protein products which, though not highly homologous to one another, display the hallmarks of the cathepsin B class of proteolytic enzymes, and therefore support the naming of these genes as cathepsin B-like.
4.2.8. Phylogenetic analysis

Phylogenetic analysis requires that the sequences to be analysed are aligned accurately. The Clustal program, version V (Higgins et al., 1992) included in the PHYLIP package version 3.5c (Felsenstein, 1993), was used to generate all alignments for phylogenetic analysis. These alignments were checked by eye to ensure that anomalous gaps had not been introduced into the protein sequences to be analysed. Once the sequences had been aligned, two distinct methods, parsimony and distance-matrix, were used to estimate the evolutionary trees from these alignments, using the PROTPARS and NEIGHBOR programs, respectively, of the PHYLIP package version 3.5c (Felsenstein, 1993).

The PROTPARS method, like all parsimony approaches, can be used to infer phylogenies directly from the character data obtained from the alignment. In essence this means scoring each character (amino acid residue) for similarity to characters in homologous positions from the other sequences. As is suggested by their names, parsimony algorithms assume that the shortest evolutionary trees are the nearest approximation to the 'true' trees. Thus these programs select the trees that minimise the total tree length (i.e. minimise the number of evolutionary steps, or transformations from one character state to another, required to explain a given data set). The PROTPARS program uses several constraints when making this comparison. First, it insists that any changes of amino acid be consistent with the genetic code. However changes between two amino acids via a third are allowed and counted as two changes. Second, the algorithm does not score synonymous changes, nucleotide changes which do not alter the amino acid sequence. This is because DNA sequence comparisons suggest that synonymous changes are considerably faster and easier to make than ones that change the amino acid sequence. The algorithm therefore assumes that such changes need not be counted.

The best way to obtain an optimal tree, defined as the shortest tree when using parsimony analysis, is to determine all the possible trees that can be generated from a given dataset and subsequently select the shortest tree. Such exhaustive methods become impractical with large datasets, since they require substantial amounts of computer processor time. Heuristic approaches are therefore used which sacrifice the
guarantee of optimality in favour of reduced computing time. The PROTPARS program is one example of many heuristic approaches which can be used to recreate a phylogenetic tree. This program recreates phylogenetic relationships by stepwise addition. Essentially this involves selecting three taxa (or sequences) to generate an initial tree. In the next step, one of the as yet unplaced taxa is added to the tree. In this process, each of the three trees that would be generated by adding this next taxon (or sequence) along one of the three branches, are evaluated and the optimal tree selected for the next round of addition. The process is then repeated for the next unplaced taxon until all the taxa have been placed on the tree. Thus, at each round of addition of a new taxon or sequence, the PROTPARS program selects the optimal tree (the shortest tree) prior to adding a new taxon. With such a method, locally optimal solutions will be obtained but there is no guarantee that the resulting tree will represent the globally optimal solution. However, the success of such a method in obtaining a globally optimal solution can be tested statistically. The bootstrap resampling technique (Efron and Gong, 1983) is one example of such a statistical test and was used to analyse some of the evolutionary trees recreated by the PROTPARS program. This technique takes random samples from the initial dataset (the alignment) and generates a new tree in each case. This process is repeated several times, and the number of times each grouping in the original tree occurs in the bootstrap trees is counted. These figures can then be used as a guide to the reliability of the clustering observed in the tree.

A distance-matrix approach was also used to infer phylogenetic relationships from the aligned sequences. When using such an approach the protein sequence alignments must first be transformed into pairwise-distances between each sequence in the alignment. This was achieved using the Dayhoff PAM 001 distance matrix (Dayhoff et al., 1978), which generates a transition probability matrix that allows prediction of the probability of changing from one amino acid to any other. These probabilities are scaled in terms of a unit, which is an expected 1% change between two amino acid sequences. The program then uses these probabilities to compute the distance between each pair of taxa. The distance calculated is therefore scaled in units of the expected fraction of amino acids changed. The phylogenetic relationships of the taxa were then recreated from this distance-matrix using the neighbor-joining algorithm (Saitou and Nei, 1987) of the NEIGHBOR program. This approach modifies the raw data of the distance-matrix
by adjusting the distances between each pair of taxa on the basis of the average divergence from all other taxa pairs. The algorithm starts by linking the least distant pair of taxa as defined by the modified matrix. Once this has been done the common node, defined by the common ancestor of these two taxa, is added to the tree but the branches are removed. This process converts the common ancestor into a terminal node on a tree of reduced size. This process is repeated, with the loss of two branches at each step, until the process is completed, when there are two nodes separated by a common branch. Though this approach is distinct from parsimony approaches, it still utilises the same principle of maximum parsimony (or minimum-evolution) when generating a tree. However, this method does not necessarily produce the minimum-evolution tree, since, like heuristic methods, it does not generate all possible trees and then select the shortest tree.

It is important to note that the phylogenetic trees generated by these two approaches have very different appearances. Both approaches recreate phylogenetic relationships by linking more closely related sequences by fewer internal nodes. However, unlike the PROTPARS program, the NEIGHBOR program also computes branch lengths based on the degree of similarity of sequences linked by a common node. Thus, the trees generated by the NEIGHBOR program have branch lengths of different sizes proportional to the degree of similarity between sequences while the PROTPARS program generates trees with all branch lengths set to the same size.

4.2.8.1. The predicted proteins encoded by cpr-3, cpr-4, cpr-5 and cpr-6 cluster with known cathepsin B enzymes

As discussed above, the primary amino acid sequence data suggested that cpr-3, cpr-4, cpr-5 and cpr-6 encoded cathepsin B-like enzymes. In order to confirm this, the nucleotide sequences from a variety of cysteine protease genes were obtained from the Genbank and EMBL databases. These sequences were identified by two searches using the STRINGSEARCH program of the GCG package, with the search terms ‘cysteine protease’ and ‘cathepsin’, respectively. Analysis of these sequences revealed that only 62 encoded the complete predicted protein product, and these were selected for initial phylogenetic analysis. In the case of genomic sequences, information supplied with the
sequences was used to remove all non-coding regions, such as 5' and 3' flanks and introns, to generate complete coding sequences. For sequences derived from cDNA, all 5' and 3' UTR sequences were removed to generate the complete coding sequences. The protein sequences for each of the coding nucleotide sequences obtained were subsequently determined using the TRANSLATE program of the GCG package. The sequences were aligned using the Clustal program version V, and phylogenetic relationships recreated using the NEIGHBOR and PROTPARS programs of the PHYLPIL package, version 3.5c. The results of these analyses are shown in Figures 4.18 and 4.19.

The NEIGHBOR program (Figure 4.18) clusters all four protein products of cpr-3, cpr-4, cpr-5 and cpr-6 with the 17 cathepsin B and B-like sequences included in the analysis, strongly supporting the assignment of these four genes as cathepsin B-like. All these cathepsin B-like sequences clustered to a single internal node representing the common cathepsin B ancestor (node A). Furthermore, this node is well separated from a second node representing the common ancestor of all the other cysteine protease enzymes analysed (node B). The degree of separation between nodes A and B suggests that cathepsin B enzymes have diverged substantially from other classes of cysteine protease. The PROTPARS program obtained two equally parsimonious trees. These two trees were virtually identical and therefore only one tree, and the part of the second tree that differed, is shown (Figure 4.19). The PROTPARS program also clusters all the cathepsin B-like sequences to one end of the unrooted tree. However, four of the C.elegans cathepsin B-like genes, cpr-1, cpr-3, cpr-4 and cpr-5, are placed very near sequences not considered as being cathepsin B-like. In particular, cpr-4 and cpr-5 are only separated by three internal nodes from a cysteine protease isolated from Dictyostelium discoideus, yet are separated from cpr-1 and cpr-3 by four internal nodes. By comparison, the NEIGHBOR program separated cpr-4 and cpr-5 from their nearest non cathepsin B-like sequence by seven internal nodes, and from cpr-1 and cpr-3 by only four internal nodes. Therefore, while the PROTPARS program does support, to some extent, the assignment of cpr-3, cpr-4, cpr-5 and cpr-6 as cathepsin B-like, it does not cluster these sequences as closely to the other cathepsin B-like sequences, or as far from non cathepsin B-like sequences as the NEIGHBOR program. The variation of the phylogenetic relationships recreated by these two programs is also manifested in other
regions of the trees. Thus, while in most cases both approaches cluster the same groups of sequences together, the relationships between the members of these groups differ between the two trees. These differences are probably accounted for by the different approaches used. As explained earlier, the PROTPARS program uses a heuristic approach to recreate the phylogenetic relationships. Such approaches do not guarantee the generation of the most parsimonious, or minimum-evolution tree. In the case of the PROTPARS program, an additive approach is used when recreating the phylogenetic relationships between sequences, taking the order of the sequences in the alignment as the order for addition to the tree. Therefore, the program can only be expected to generate locally optimal relationships, and these may differ depending on the order the sequences are added. In essence, the program is unable to take a global approach to recreating the phylogenetic relationships and can only generate the most optimal relationships given the current situation (the input order of the sequences). In such situations bootstrap analysis is useful in determining the reliability of trees inferred by such programs. However, the large size of the dataset used negated this approach as it would have proved to be too lengthy an operation.

In contrast, the NEIGHBOR approach first generates a distance matrix, and therefore organises the input sequences on the basis of similarity. Though these distances are then used to recreate phylogenetic relationships using an additive approach, the sequences are added in a logical manner, starting with the most similar sequences first. In this respect the NEIGHBOR program can recreate phylogenetic relationships using a more global approach and is therefore a more appropriate algorithm to use for such a large dataset. Despite the shortcomings of the PROTPARS approach, it was still able to support results obtained using the NEIGHBOR program. Together, these data strongly support the classification of the genes cpr-3, cpr-4, cpr-5 and cpr-6 as cathepsin B-like, and indicates that this class of cysteine protease enzyme is highly diverged from other classes of cysteine protease.
4.2.8.2. Phylogenetic comparison of the proteins encoded by cpr-3, cpr-4, cpr-5 and cpr-6 with cathepsin B-like sequences in the database

A comparison of the percentage similarities and identities of the prepro- forms and predicted mature forms of the enzyme products of cpr-3, cpr-4, cpr-5 and cpr-6 to cpr-1 and human cathepsin B, suggested that the cathepsin B-like genes of C.elegans were as diverged from one another as from human cathepsin B, with the exception of cpr-4 and cpr-5 (Table 4.4). The NEIGHBOR and PROTPARS programs were therefore used to infer the phylogenetic relationships of the 21 cathepsin B-like sequences which clustered together in the previous phylogenetic analysis (Section 4.2.8.1).

The 21 cathepsin B-like sequences were realigned using the CLUSTAL program, version V. The NEIGHBOR and PROTPARS programs were used to recreate the phylogenetic relationships between these sequences as described previously. Bootstrap resampling was also performed with the PROTPARS parsimony analysis. A majority-rule consensus tree was obtained from the 20 trees generated by the bootstrap analysis using the CONSENSUS program of the PHYLIP package, version 3.5c (Felsenstein, 1993). These two approaches generated trees of very similar topology (Figure 4.20), with only three minor differences between the two trees. These differences were essentially in the relationship of AC-4 with the other H.contortus cathepsin B-like genes, cpr-6 with cpr-4 and cpr-5, and human cathepsin B with the other mammalian cathepsin B-like genes. The very similar topologies obtained from the NEIGHBOR and PROTPARS programs used strongly suggest that the overall phylogenetic relationships inferred by these two approaches are reliable.

The cathepsin B-like sequences can be divided into several groups based on taxonomic differences; the protist sequences (Leishmania mexicana), the plant sequences (Nicotiana rustica and Triticum aestivum), the vertebrate sequences (Homo sapiens, Mus musculus, Bos taurus and Gallus gallus), the insect sequences (Sarcophaga peregrina), the trematode sequences (Schistosoma japonicum and Schistosoma mansoni) and the nematode sequences (C.elegans, H.contortus and O.ostertagi). Bootstrap analysis indicated that sequences from the same taxonomic group were in most cases reliably clustered together (i.e. 100% of the bootstraps
recreated these clusters). However the notable exception to this proved to be with the *C.elegans* sequences. Thus, while *cpr-1* and *cpr-3* were reliably clustered together, as were *cpr-4* and *cpr-5*, neither of these two groups were reliably clustered to each other or to *cpr-6*. This is in contrast to the sequences of *H.contortus* and *O.ostertagi* which were all reliably clustered together. The unreliable clustering of these sequences probably accounts for the discrepancy in the relationship of *cpr-6* to *cpr-4* and *cpr-5* observed between the two trees. The inability of bootstrap analysis to reliably group the cathepsin B-like sequences of *C.elegans* may be a result of the high level of divergence observed between these sequences. The phylogenetic relationships inferred using the NEIGHBOR and PROTPARS program suggest that there are three groups of cathepsin B-like enzymes in *C.elegans*. These three groups are represented by *cpr-1* and *cpr-3*, *cpr-4* and *cpr-5*, and lastly, *cpr-6*. The NEIGHBOR program generates long branch lengths between these three groups, suggesting that they are highly diverged from one another. The degree of divergence observed between the three different groups of cathepsin B-like amino acid sequences of *C.elegans* is far greater than the interspecies divergence observed for human cathepsin B and its homologue in the mouse, the chicken and the cow, suggesting that the cathepsin B-like gene family of *C.elegans* may have diverged early in evolution. For the three groups of predicted proteins encoded by the *C.elegans* cathepsin B-like genes, any one member of a group is almost as diverged from the members of other groups as it is from the vertebrate cathepsin B sequences, but is significantly less diverged from members of the same group. This suggests that these three groups may have arisen from gene duplication events which occurred around the time of separation of the nematode and vertebrate ancestral lineages, but that more recent duplication events may have subsequently generated the members comprising each group. However, an important caveat to this observation is that one must assume an equal evolutionary rate when making such comparisons.

4.2.8.3. Gene architecture as a gauge of evolutionary divergence

As mentioned in the previous section, it is necessary to assume that evolution has occurred at an equal rate, when using the degree of divergence between related sequences to infer how ancient the gene duplication events were which gave rise to such
sequences. When comparing orthologous genes (genes related by a common ancestor which perform analogous roles in different species), this assumption may be correct. However, such an assumption may not be valid when making comparisons between paralogous sequences. Paralogous sequences represent genes that, while ultimately having a common ancestor, are derived from a gene duplication event rather than a speciation event. After duplication, the paralogues may have assumed different physiological or enzymatic roles and therefore may have been subjected to different evolutionary pressures. These differences in evolutionary pressure may increase the rate at which the coding sequences of two paralogous genes evolve, resulting in highly diverged genes which have actually arisen from more recent duplication events than would be predicted from the degree of divergence of their amino acid sequences. In such circumstances, it would be useful to select a feature which is unlikely to be affected by differing evolutionary pressures to the extent that coding sequences are. In this respect introns may represent a suitable feature, since evolutionary pressures are unlikely to alter the rate at which gene architecture is altered with time. Furthermore, the mutation events which cause intron loss, gain or displacement and lead to altered gene architectures are less likely to generate functional proteins than substitution mutations that give rise to diverged predicted amino acid sequences. Therefore the frequency with which these mutations become fixed during evolution will be somewhat less than substitution mutations that result in changes in the amino acid sequence. Therefore, it is possible to envisage a situation where more recently duplicated paralogous genes may encode highly diverged protein products but share very similar gene architecture. In such situations, the gene architecture will provide important clues as to how recently the genes diverged. For this reason, I decided to study the relative position of introns in cathepsin B and cathepsin B-like genes from different species.

The intron positions of eight cathepsin B and B-like genes were determined from their gene architectures. The predicted amino acid sequences of these genes, and the relative position of their introns, are shown in Figure 4.21. Before discussing the evolutionary information that can be inferred from the intron positions of the eight genes analysed, a brief overview of the patterns of intron distribution in these genes will be mentioned. A comparison of the relative intron positions between the eight genes reveals that a total of six intron sites are conserved between the genes (Figure
4.21, A-F). These sites are not only conserved with respect to the amino acid sequence but also with respect to the exact nucleotide residue at which intron / exon boundaries occur within codons. This level of conservation strongly suggests that these eight genes are derived from a common ancestor. However, there is also evidence to suggest that intron loss, gain and displacement has occurred during the evolution of these genes. If one looks at the intron positions defined by the 11 introns of the *H. contortus* AC-2 gene and *O. ostertagi* CP-1 gene, it becomes apparent that there are several incidences where no introns are found at the equivalent position in the other genes (Figure 4.21, A-H), suggestive of intron loss. A clear example of this is seen for *cpr-6* and mouse cathepsin B at sites E and F respectively. In this example, *cpr-6* has maintained the intron at position E and lost the intron at position F, while the opposite has occurred with mouse cathepsin B. There are also some incidences where the introns of the *C. elegans* cathepsin B-like genes and the mouse cathepsin B gene are found in the vicinity of those sites defined by AC-2 and CP-1, but not at exactly the same positions (Figure 4.21, A, I, J and K), suggesting that intron displacement has occurred with some of these genes presumably as a result of complementary mutation events occurring at each end of the intron. Finally, there are two cases where introns are found in close proximity to each other, but outwith those sites defined by AC-2 and CP-1 (Figure 4.21, L and M). Such positions may have arisen as a result of intron loss in CP-1 and AC-2 followed by intron displacement in the genes containing these introns. Alternatively, a simpler explanation may be that introns were inserted at these sites, after divergence from the *H. contortus* and *O. ostertagi* ancestral lineages. Such complex patterns of intron position have also been observed for other gene families. Indeed, the hemoglobin family of diverged nematode globins shows similar types of variation in intron position which have also been explained by models of intron loss, gain and displacement (Blaxter, 1993).

For evolutionary analysis, the AC-2 and CP-1 genes of the parasitic nematode species *H. contortus* and *O. ostertagi* provide a useful baseline for comparison with the other genes. Phylogenetic analysis of the protein products of AC-2 and CP-1 (Figure 4.20) suggest that these two genes are very closely related and this is reflected by their very similar gene architectures. Thus, both genes possess 11 introns and the position of all these introns are not only conserved with respect to the amino acid sequence but also to the exact nucleotides at which intron / exon boundaries occur within codons (Figure
Furthermore, these two genes define all the common intron sites for the genes analysed since the *C. elegans* cathepsin B-like genes and the mouse cathepsin B gene only share common intron positions at sites already determined by AC-2 and CP-1 (Figure 4.21, A-F). Accordingly, the 11 intron positions defined by AC-2 and CP-1 were used as the basis of comparison for the five *C. elegans* cathepsin B-like genes and the mouse cathepsin B gene included in this analysis.

Alignment of the amino acid sequences of the eight genes revealed that only six of the 11 common intron positions defined by AC-2 and CP-1 are shared with the five *C. elegans* cathepsin B-like genes and the mouse cathepsin B gene (Figure 4.21, A-F). This suggests that the *C. elegans* and mouse genes are substantially more diverged from AC-2 and CP-1 than these two genes are from one another. This conclusion is also supported by the variation in intron number of the *C. elegans* and mouse genes. All six of these genes possess substantially fewer introns in their coding region than AC-2 and CP-1, with *cpr*-6 and mouse cathepsin B having the most (seven introns) and *cpr*-4 the least (one intron). Furthermore, none of the *C. elegans* and mouse genes each share more than four common intron positions with AC-2 and CP-1. Thus, while mouse cathepsin B and *cpr*-6 both possess seven introns in their coding regions, each shares only four and three common intron positions, respectively, with those sites defined by AC-2 and CP-1. Together these data clearly suggest that the five *C. elegans* cathepsin B-like genes and the mouse cathepsin B gene are all much more diverged from AC-2 and CP-1 than these two genes are from one another. This conclusion was also drawn from the phylogenetic analysis involving the predicted proteins of these eight genes (Figure 4.20) and therefore supports the efficacy of using gene architecture, as defined by intron position, as a means of gauging evolutionary divergence.

A comparison of common intron positions between the eight genes (Figure 4.21, A-F) strongly suggests that most of the *C. elegans* cathepsin B-like genes are as diverged from one another as they are from mouse cathepsin B. None of the *C. elegans* cathepsin B-like genes share more common intron positions with one another than they do with mouse cathepsin B. In fact, no more than one common intron position is shared between any of the *C. elegans* and mouse genes. For example *cpr*-6 shares only one common site with *cpr*-1, *cpr*-4 and the mouse cathepsin B gene, and no common sites with *cpr*-3 or
cpr-5. This level of divergence suggests that the *C. elegans* cathepsin B-like genes have arisen from ancient gene duplication events.

Phylogenetic analysis of the predicted proteins encoded by the five cathepsin B-like genes of *C. elegans* (Figure, 4.20) clusters these sequences into three groups and suggests that each group may have arisen from gene duplication events which occurred as early as the time of separation of the nematode and vertebrate ancestral lineages, but that the members within each group may have arisen from more recent duplication events. The three groups comprise; *cpr-1* and *cpr-3*, *cpr-4* and *cpr-5*, and *cpr-6*. In some cases, the results obtained from phylogenetic analysis of the predicted encoded proteins agree with those obtained from analysing the gene architectures. For example, phylogenetic analysis of the predicted encoded proteins suggests that the group defined by *cpr-6* is highly diverged from the other two groups. This is supported by the gene architectures, as defined by intron position and number. *cpr-6* possesses seven introns while the members of the other two groups possess only between one and three introns. Furthermore, *cpr-6* does not share more than one intron position in common with any of the four other *C. elegans* cathepsin B-like genes comprising the other two groups. Phylogenetic analysis suggests that *cpr-1* and *cpr-3* may have arisen from more recent gene duplication events than those that generated the three groups. This is supported, at least in part, by comparison of the gene architectures of these two genes. *cpr-1* and *cpr-3* have a similar number of introns, two and three introns respectively. Furthermore, intron I of *cpr-1* is situated in close proximity to intron I of *cpr-3* (Figure 4.21, A), and intron II of *cpr-1* is situated at an identical position to intron II of *cpr-3* (Figure 4.21, D). Though these two genes do possess some similarities in their gene architectures (suggestive of a more recent gene duplication event) their gene architectures are sufficiently diverged to suggest that this duplication event may have occurred soon after the events predicted to have generated the three groups. This is supported by phylogenetic analysis using the NEIGHBOR program which, links the products of these two genes by a common node but separates them with long branch lengths (Figure 20C).

For *cpr-4* and *cpr-5*, there is no such correlation between the similarity of their predicted proteins and their gene architectures. The putative enzymes encoded by *cpr-4* and *cpr-5* show higher levels of homology to each other than the predicted proteins encoded by any other pair of *C. elegans* cathepsin B-like genes (Table 4.4) and this is
reflected by the results obtained from phylogenetic analysis (Figure 4.20C). However, their gene architectures, as defined by intron number and position, are very different. *cpr-4* possesses a single intron and the position of this intron is not conserved with either of the two introns of *cpr-5*, nor are the introns of these two genes found in close proximity to one another. This suggests that *cpr-4* and *cpr-5* may have arisen from a much more ancient gene duplication event than expected from the similarity of their protein sequences. Accordingly, the coding regions of these two genes may have evolved either in parallel or at a much slower rate after gene duplication than the other *C. elegans* genes, suggesting that these two genes may encode enzymes which are required for similar biological functions.

4.3. Discussion

4.3.1. A cathepsin B-like multigene family in *C. elegans*

The results presented in this chapter indicate that *C. elegans* possesses a cathepsin B-like cysteine protease multigene family comprising the four genes *cpr-3*, *cpr-4*, *cpr-5* and *cpr-6* (Chapter 3) and the previously isolated cathepsin B-like gene, *cpr-1* (Ray and McKerrow, 1992). Sequence analysis indicates that the putative cathepsin B-like enzymes encoded by *cpr-3*, *cpr-4*, *cpr-5* and *cpr-6* possess all the features expected for this class of enzyme (Section 4.2.7). Furthermore, phylogenetic analysis clusters the predicted enzyme products of these four genes with other cathepsin B sequences in the database and away from other cysteine protease classes (Section 4.2.8.1).

4.3.2. The Implications of isolating a cathepsin B-like multigene family from *C. elegans*

Until now, cathepsin B-like multigene families had only been reported in parasitic nematode and trematode species (Chapter 1, Section 1.9) leading to suggestions that amplification of this gene family may have been important for the parasitism of these species (Pratt et al., 1992a). However, the presence of a cathepsin B-like multigene
family in the free living nematode *C. elegans*, indicates that such gene families are not unique to parasitic nematode and trematode species *per se*.

4.3.2.1. *C. elegans*, *H. contortus* and *O. ostertagi* may possess more, as yet, unidentified cathepsin B-like genes

Phylogenetic comparisons using the NEIGHBOR and PROTPARS programs of the PHYLIP package, reveal interesting differences in the relationships of the predicted products of the cathepsin B-like genes of *C. elegans*, both to each other and to the products of cathepsin B, and cathepsin B-like genes from other organisms (Section 4.2.8.2). In particular, the products of the cathepsin B-like genes of *C. elegans* are much more diverged from each other than those isolated from *H. contortus* and *O. ostertagi*. This may be explained by the methods used to isolate these genes. The *H. contortus* cathepsin B-like genes AC-2, AC-3 and AC-4 and the *O. ostertagi* gene CP-1 were all isolated by homology to the previously isolated gene AC-1. AC-1 was initially isolated by screening a cDNA expression library using antisera generated from proteins purified from anticoagulant extracts of *H. contortus* (Cox et al., 1990). AC-2 was subsequently isolated from an *H. contortus* genomic DNA library using AC-1 as a probe (Pratt et al., 1990). These two sequences share 97% nucleotide sequence identity and 98% amino acid sequence identity in their coding regions. Though it is unclear whether these sequences represent distinct genes, or population polymorphisms, the authors argue that the number of differences is more consistent with their being distinct genes. AC-3 was initially identified by low stringency hybridisation to the genomic clone containing AC-2 using a 40 nucleotide oligomer corresponding to the 5' end of the AC-1 cDNA, also present in AC-2 (Pratt et al., 1992a). Subsequent sequence analysis allowed the design of a non-degenerate oligonucleotide primer which was used, in conjunction with an oligo (dT) primer, to amplify sequences from *H. contortus* cDNA, and yielded cDNA clones for AC-3 and AC-4. CP-1 and CP-2 (a partial genomic sequence) were isolated by screening an *O. ostertagi* genomic DNA library with the AC-1 cDNA of *H. contortus* (Pratt et al., 1992b). Such homology based methods inevitably bias the screen by failing to detect divergent members.

In contrast, the method by which the genes reported here were identified has no such bias since the cDNA clones used to isolate *cpr*-3, *cpr*-4, *cpr*-5 and *cpr*-6 were
originally identified by partial sequencing of randomly selected clones (Waterston et al., 1992). Indeed, the library from which the cDNA clones of cpr-3, cpr-4, cpr-5 and cpr-6 were obtained, may have biased against the isolation of highly homologous clones because the library was partially normalised prior to isolation of the cDNA clones. This normalisation was achieved by screening a lambda SHLX2 C.elegans cDNA library with cDNA clones and selecting only clones that did not hybridise. These clones were then added to those used for probing and the process repeated several times (Palazzolo et al., 1990).

The differences in the degree of similarity of the genes cloned by these two approaches may therefore be due to different biases imposed by the methods used for their isolation. It is therefore possible that these nematode species may possess more members of the cathepsin B-like gene family than those so far identified. If this is the case, one might expect the cathepsin B-like multigene families of nematode species to be organised into several distinct highly diverged groups (as suggested from the C.elegans data), with each group comprising several highly homologous members (as suggested by the H.contortus and O.ostertagi data).

4.3.2.2. The cathepsin B-like multigene families of free living and parasitic nematode species may have arisen from different gene duplication events

The greater degree of divergence observed between most of the C.elegans cathepsin B-like genes than that observed between the H.contortus and O.ostertagi cathepsin B-like genes may reflect a genuine difference between these two species, not just different biases in the methods used to isolate these genes. In such a situation some of the members of the cathepsin B-like multigene families of H.contortus and O.ostertagi may have arisen from more recent duplication events than those of C.elegans, as a result of evolutionary pressures unique to these parasitic nematode species.

Most of the cathepsin B-like genes of C.elegans are sufficiently diverged to suggest that they arose from duplication events which occurred early in evolution, possibly as early as the time of separation of the nematode and vertebrate ancestral lineages (Sections 4.2.8.2 and 4.2.8.3). Since H.contortus, O.ostertagi and C.elegans do not represent highly diverged taxa, being of the same Subclass (Secerentea), the ancestral
lineages of these nematode species probably separated after the ancient gene duplication events predicted to have generated the three groups of cathepsin B-like genes in *C. elegans*. Accordingly, these three nematode species are likely to possess cathepsin B-like genes which arose from the same duplication events that generated the three groups of cathepsin B-like genes identified in *C. elegans*. However, subsequent duplication events in the parasitic nematode species, after separation of the *C. elegans* ancestral lineage, may have generated additional members of this multigene family in these parasitic nematode species. This scenario predicts that *H. contortus* and *O. ostertagi* may possess more cathepsin B-like genes than *C. elegans* and that these additional genes will be very closely related. There are several pieces of evidence to suggest that this is indeed the case. There is circumstantial evidence to suggest that *H. contortus* may possess more cathepsin B-like genes than *C. elegans*. At least six additional cathepsin B-like genes have been identified from *H. contortus* (D. Knox, pers. comm.) indicating that, in this parasitic nematode species, the gene family comprises at least 11 members. In contrast, only one more cathepsin B-like cDNA clone (distinct from the cathepsin B-like genes isolated to date) has been isolated by partial sequencing of randomly isolated cDNA clones from *C. elegans* (McCombie et al., 1992) and no additional cathepsin B-like genes have been identified in the 22.5 Mb of *C. elegans* genomic DNA sequenced to date. There is also evidence to suggest that the *H. contortus* and *O. ostertagi* cathepsin B-like genes characterised to date have arisen from more recent duplication events than the *C. elegans* cathepsin B-like genes. The gene architectures (as defined by intron number and position) of AC-2 and CP-1 from *H. contortus* and *O. ostertagi* are identical, both possessing 11 introns at identical positions. Partial sequencing of the 3′ end of the *O. ostertagi* CP-3 gene has identified four introns which are in identical positions to introns 9 - 11 of CP-1 and AC-2 (Pratt et al., 1992b). In addition, partial sequencing of the 5′ end of the *H. contortus* AC-3 gene has identified four introns which are in identical positions to introns 1 - 4 of AC-2 and CP-1 (Pratt et al., 1992a). These results suggest that the four genes, AC-2 and AC-3 from *H. contortus* and CP-1 and CP-3 from *O. ostertagi*, may all have introns in identical positions and therefore share very similar gene architectures. These observations suggest that these genes have arisen from a much more recent duplication event than the *C. elegans* cathepsin B-like genes. One would expect the ancestor of these duplication events to be one of the genes generated by the
same ancient duplication event predicted to have generated the three groups of *C. elegans* cathepsin B-like genes. Accordingly, one might expect to find at least one cathepsin B-like gene in *C. elegans* that is significantly more closely related to the cathepsin B-like genes isolated from *H. contortus* and *O. ostertagi*, which would represent the orthologue of the gene amplified in these parasitic nematode species. Such a gene has not as yet been identified in *C. elegans* since none of the cathepsin B-like genes isolated from *C. elegans* to date possess gene architectures or encode enzymes which are significantly more similar to the *H. contortus* and *O. ostertagi* genes. Therefore, there may be at least one, as yet uncharacterised, cathepsin B-like gene in the genome of *C. elegans* which will show more homology to the AC and CP cathepsin B-like genes of *H. contortus* and *O. ostertagi* than the genes isolated to date.

If the members of the cathepsin B-like multigene families of *H. contortus* and *O. ostertagi* isolated to date have arisen in the manner suggested above, the enzymes encoded by these genes may have important roles required for parasitism by these two species. Accordingly, although the isolation of a cathepsin B-like multigene family from the free living nematode *C. elegans* indicates that such multigene families are not unique to parasitic nematode species *per se*; it may not disprove the specific function of some cathepsin B-like genes in parasitic processes.

### 4.3.2.3. Cathepsin B-like multigene families may be present in other metazoans

In order to argue the presence of more diverged cathepsin B-like genes in the genomes of higher eukaryotes, or indeed of *H. contortus* or *O. ostertagi*, one must identify examples of genes which encode cathepsin B-like enzymes that are sufficiently diverged not to cross hybridise with one another. The COMPARE and GAP analyses performed between the genomic sequences of *cpr*-3, *cpr*-4, *cpr*-5 and *cpr*-6 indicate these genes are highly diverged (Section 4.2.5), with only short regions of homology occurring between the genes in the regions which encode the catalytic residues, C, H and N. This suggests that these genes are sufficiently diverged such that detection of one, using another as a hybridisation probe, would be extremely difficult, or impossible. The degree of divergence therefore suggests that it is possible that more diverged cathepsin B-like genes might not be detected with a homology based screen, and therefore that
*H. contortus* and *O. ostertagi* may possess yet more members of this gene family than those isolated to date. This observation has wider implications since additional diverged cathepsin B-like genes may also exist in other metazoans in which only single cathepsin B genes have so far been found.

For example, there is a large amount of circumstantial evidence to suggest that the human and mouse cathepsin B genes are single copy. The evidence gathered to date which suggests that the human and mouse cathepsin B genes are single copy, comes from data obtained during library screens to obtain cDNA clones and genomic clones for these genes. Thus, the first cDNA clones of human and mouse cathepsin B to be identified from lambda gt11 library screens were all overlapping (Chan et al., 1986). The authors concluded that each of these genes are single copy. This conclusion was supported by the simple hybridisation patterns obtained by hybridising Southern blots of restriction endonuclease digested human genomic DNA with a human cathepsin B cDNA clone (Fong et al., 1986). The genes encoding human and mouse cathepsin B have subsequently been cloned (Gong et al., 1993; Ferrara et al., 1990; Qian et al., 1991). More recently, a number of human cathepsin B cDNA clones have been cloned, sequenced (Gong et al., 1993; Cao et al., 1994; Tam et al., 1994) and compared to the human cathepsin B gene sequence. These comparisons all suggest that the cDNA clones isolated are all derived from the same human cathepsin B gene. These and other studies have resulted in the current opinion that human and mouse cathepsin B (and therefore probably also the rat and bovine cathepsin B homologues) are present as single copies in their respective genomes. This does not dispute the presence of additional, diverged cathepsin B-like genes within the human and mouse genomes, since such genes may have remained undetected by the homology based screens that have been used to obtain evidence for the single copy nature of the human and mouse cathepsin B genes. Thus there may also be additional, diverged cathepsin B-like genes present in the human and mouse genomes which have not as yet been detected. This would be very similar to the situation in *C. elegans*, where the results suggest that each of the cathepsin B-like genes, *cpr-3, cpr-4, cpr-5* and *cpr-6*, are single copy (Chapter 3) and are sufficiently diverged such that detection of one using another as a hybridisation probe may be difficult or impossible.
Alternatively, the existence of a cathepsin B-like multigene family within nematode species may reflect functions specific to nematode biology. Accordingly, vertebrates may not possess additional, diverged cathepsin B-like enzymes. If this proves to be the case, cathepsin B-like enzymes may prove to be very useful targets for anthelminthic agents in the treatment of parasitic nematode diseases.

4.3.3. The cathepsin B-like genes of *C. elegans* may encode enzymes with different biological functions

As stated above, the predicted amino acid sequences of *cpr*-3, *cpr*-4, *cpr*-5 and *cpr*-6 strongly suggest that these genes encode cathepsin B-like enzymes. However, these putative enzymes are much more diverged from one another than those encoded by the genes, AC-1, AC-2, AC-3, AC-4 and AC-5 of *H. contortus*. The five cathepsin B-like genes of *C. elegans* can be divided into three groups, on the basis of the similarity of their predicted enzyme products (see Section 4.2.8.2). Phylogenetic analysis reveals that each of these three groups are almost as diverged from one another as they are from the vertebrate cathepsin B enzymes. Such a high degree of intraspecies divergence suggests that the three groups of enzymes may have evolved independently to assume distinct biological functions.

4.3.3.1. *cpr*-1, *cpr*-3, *cpr*-4, *cpr*-5 and *cpr*-6 may encode enzymes with different specificities and activities

Recombinant rat cathepsin B has been expressed in yeast and active enzyme has accumulated in the media under optimal culture conditions (Lee *et al.*, 1990). Detailed kinetic analysis of the purified recombinant protein obtained indicates that enzymatically it is essentially identical to rat liver cathepsin B (Hasnain *et al.*, 1992). This system has allowed site-directed mutagenesis approaches to be used to study the effect of altering several residues previously implicated as being important in determining the activity of cathepsin B. These studies (Hasnain *et al.*, 1992) have identified two regions outside the catalytic residues C, H and N, which affect the substrate specificity and activity of rat cathepsin B. Cathepsin B appears to be unusual amongst other members of the cysteine
protease group of proteolytic enzymes in its ability to function as both an endopeptidase and a peptidyl dipetidase (Aronson and Barrett, 1978). Two adjacent histidine residues at position 110 and 111 of rat cathepsin B are thought to play an important role in this peptidyl dipetidase activity (exopeptidase activity). Mutation of one of the histidine residues, His(111), to glutamine results in a much reduced exopeptidase activity and increased endopeptidase activity with respect to wild-type rat cathepsin B, providing strong evidence that the dihistidine repeat plays an important role in exopeptidase activity. In the second region, mutation of the glutamic acid residue (position 245 in mature rat cathepsin B) to glutamine or alanine alters the substrate specificity of the recombinant enzyme. This residue, Glu(245), is thought to be responsible for vertebrate cathepsin B’s characteristic activity against synthetic substrates possessing two basic amino acid residues. Accordingly, mutation of this residue in recombinant rat cathepsin B drastically reduces the rate at which this enzyme cleaves the synthetic substrate Z-Arg-Arg-pNA, compared to wild type.

The equivalent residues for human cathepsin B are shown in bold and boxed (Section 4.2.7, Figure 4.8). A comparison of these residues to residues at the equivalent position in cpr-1, cpr-3, cpr-4, cpr-5 and cpr-6 reveals interesting differences between these enzymes at both these sites. Only cpr-6 contains the adjacent histidine residues involved in exopeptidase activity. cpr-4 and cpr-5 both share the same non-conservative substitutions of glutamic acid and threonine at this site while cpr-1 and cpr-3 share a common seven amino acid deletion of this region. This suggests that cpr-1, cpr-3, cpr-4 and cpr-5 may have reduced exopeptidase activity with respect to both cpr-6 and human cathepsin B. The predicted enzymes encoded by cpr-1, cpr-3, cpr-4, cpr-5 and cpr-6 can be grouped into three classes on the basis of: presence of the dihistidine repeat, substitution of the dihistidine repeat, or deletion of the dihistidine repeat. It is interesting to note that these three groups correlate with the three groups defined by phylogenetic analysis (Section 4.2.8.2). At the second site, equivalent to Glu(245) of rat cathepsin B, none of the C.elegans cathepsin B-like enzymes possess this glutamic acid residue, suggesting that these enzymes may not possess significant activities against synthetic substrates with two basic amino acid residues. Furthermore, the substitutions made at this site are not conservative. These data suggest the cathepsin B-like enzymes isolated
from *C. elegans* may have different specificities and activities not only from mammalian cathepsin B, but also from each other.

### 4.3.3.2. The enzymes encoded by *cpr-3, cpr-4, cpr-5* and *cpr-6* may be subjected to different types of post-translational modification

Only *cpr-3, cpr-4* and *cpr-6* encode products with potential N-glycosylation sites (Asn-X-Ser/Thr) and their positions are not conserved. Studies with mammalian cells have demonstrated that glycosylation is an integral part of lysosomal trafficking, with mannose-6-phosphate being the recognition signal for transport of soluble proteins to the lysosomes (von Figura and Hasilik, 1986). Briefly, N-linked glycosylation of certain asparagine residues of proteins occurs in the lumen of the endoplasmic reticulum and these glycosylated proteins are subsequently transported to the Golgi complex. In the Golgi complex, mannose residues of lysosomal proteins are phosphorylated. The specificity of this phosphorylation step is thought to be generated by a signal within the lysosomal protein. This signal is sensitive to treatment with heat, SDS and trypsin (Lang *et al.*, 1984) and is therefore thought to be dependent on the tertiary rather than primary structure of the lysosomal protein. Thus, the generation of a mannose-6-phosphate lysosomal targeting signal in the Golgi complex requires the initial presence of a glycosylated asparagine residue. Treatment with tunicamycin, which inhibits protein glycosylation, results in most cell lines secreting proteins which are normally directed to the lysosomes (Rosenfeld *et al.*, 1982). These data suggest that the gene products of *cpr-1* and *cpr-5*, which both lack potential N-glycosylation sites, may be secreted as opposed to being directed to the lysosomes. Alternatively, it is possible that the enzymes encoded by these two genes may be directed to the lysosomes by a mannose-6-phosphate independent pathway. There is evidence to suggest that such carbohydrate-independent mechanisms do exist. This conclusion has been based primarily on work with hepatocytes from I-cell diseased patients (Kornfeld, 1986). I-cell disease is a disorder which affects mannose-6-phosphate biosynthesis. In some patients suffering from this disease, lysosomal proteins are still delivered to the lysosomes, despite the absence of a mannose-6-phosphate marker. Other studies, with HepG2 cultured cells treated with
tunicamycin, have also reported targeting of lysosomal proteins to the lysosomes (Rijnboutt et al., 1991).

Phylogenetic analysis clustered the five cathepsin B-like enzymes of *C. elegans* into three groups (Section 4.2.8.2); group 1, *cpr-1* and *cpr-3*; group 2, *cpr-4* and *cpr-5*; group 3, *cpr-6*. Only one enzyme in each group possesses a potential N-glycosylation signal. This suggests that the more similar enzymes within a group may have functionally distinct roles, possibly with one being secreted and one being retained within the lysosomes of the cells which express them.

Finally, both *cpr-3* and *cpr-6* encode C-terminal extensions (31 and 19 amino acids respectively) from the site of cleavage of the human C-terminal peptide (Chan et al., 1986). In contrast, the predicted protein sequences of *cpr-1*, *cpr-4* and *cpr-5* all end in the region of the predicted site of cleavage of the human C-terminal peptide, suggesting that these enzymes do not possess such a peptide and therefore lack this aspect of processing. It is not known whether the C-terminal extensions encoded by *cpr-3* and *cpr-6* are present in the mature enzymes, as previously reported for *Trypanosoma cruzi* (Cazzulo et al., 1992), or represent much larger C-terminal peptides which are cleaved. Though the role of these extensions is unknown, their presence is likely to reflect differences in either enzyme activity or substrate specificity, or differences in post-translational processing.

Interestingly, a single intron is present in the region of each of the *cpr-3* and *cpr-6* genes which encode the C-terminal extensions. The introns-early hypothesis proposes that spliceosomal introns predate the divergence of prokaryotes and eukaryotes and that each exon encodes some 'functional unit' of a protein (Doolittle and Stoltzfus, 1993). This hypothesis also suggests that novel proteins can be generated by shuffling these exons around during evolution. Therefore, the presence of introns in the C-terminal extensions of *cpr-3* and *cpr-6*, may define some additional functional domain present in these two enzymes.

### 4.3.4. Regulation of expression of *cpr-3*, *cpr-4*, *cpr-5* and *cpr-6*

The 5′ RACE experiments generated RACE clones for each gene whose 5′ ends cluster to a single region of the 5′ flank of each gene (Section 4.2.6), suggesting that
each gene is under the control of a single promoter region. Northern analysis (Chapter
5, Section 5.2.2) later revealed that the predominant transcripts from each of these four
genes are of a single size species. Furthermore, no evidence for alternative splicing was
observed for these four genes within the limits of resolution and detection of Northern
analysis. This is in complete contrast to the human and mouse cathepsin B genes, where
the data suggest that these two genes are both transcribed from multiple promoters and
that their transcripts also undergo alternative splicing.

Northern analysis of RNA isolated from several normal murine tissues, as well as
murine tumours, have revealed multiple mRNA transcript species for mouse cathepsin B
(Qian et al., 1989; Moin et al., 1989). Similar experiments with normal human tissues
and human tumours have revealed a similar situation for human cathepsin B (Page et al.,
1992; Gong et al., 1993; Tam et al., 1994). The cloning and characterisation of human
cathepsin B (Gong et al., 1993) and mouse cathepsin B (Ferrara et al., 1990; Qian et al.,
1991) allows the comparison of cDNA clones, generated from the different mRNA
species of these two genes, with the appropriate cloned gene. As a result of this, a
number of cDNA clones derived from different sized transcripts from the human
cathepsin B gene have been cloned and sequenced (Gong et al., 1993). Comparison of
these transcripts to cloned human cathepsin B suggests that the different transcripts
arose by alternative splicing of a single gene. Five different species of transcript were
identified. Four of these contain identical coding regions, differing only in their 5’ and 3’
UTRs. The fifth lacks the signal prepeptide and a portion of the prorregion, as a result of
skipping exons two and three. However, this transcript species was only identified in
human breast and colon carcinomas and in a human melanoma, not in normal tissues, and
may therefore represent an aberrant transcript. The identification of at least four
transcript species with identical coding regions but varied 5’ and 3’ ends suggest that
human cathepsin B expression may be regulated in part at the level of RNA processing.
Interestingly, two new exons of the human cathepsin B gene have recently been
identified in the 5’ UTR between exons two and three, and have been named 2a and 2b
(Berquin, I et al., 1995). Reverse Transcription-Polymerase Chain Reaction (RT-PCR)
and primer extension assays performed with matched normal / tumour samples have
demonstrated that cathepsin B mRNA species vary both between different cell types and
between different differentiation states, and that there is a trend towards preferential
expression of exon 2a by tumours. These results suggest that RNA processing may perform an important role in the regulation of human cathepsin B in different tissues.

The same RT-PCR and primer extension assays (Berquin, I et al., 1995) have also demonstrated that transcription of human cathepsin B is initiated from more than one promoter region. This may allow for differential regulation of human cathepsin B expression from different promoters. A RACE approach has also been used to identify at least three different leader sequences for murine cathepsin B (Rhaissi et al., 1993). By comparing the nucleotide sequences of the RACE products to the 5' flank region of mouse cathepsin B, the authors confirmed that the RACE products were all derived from the same transcriptional unit but that transcription was being initiated at more than one promoter region. This suggests a similar mode of regulation for mouse cathepsin B as for human cathepsin B.

The added level of complexity in the regulation of human and mouse cathepsin B expression may reflect the greater complexity of these vertebrate species when compared to C.elegans. Thus, cathepsin B expression in vertebrates may have to be more tightly regulated than cathepsin B-like expression in C.elegans. Alternatively, the differences in the regulation of human and mouse cathepsin B gene expression, when compared to the regulation of C.elegans cpr-3, cpr-4, cpr-5 and cpr-6 cathepsin B-like gene expression, may reflect different approaches to ensuring appropriate expression of cathepsin B. Thus vertebrates may utilise a complex regulatory system to differentially regulate expression of a single cathepsin B gene in different tissues while C.elegans may use several genes which can each be regulated independently from one another to ensure expression of the appropriate gene, or combinations of genes, in the appropriate tissue and/or at the appropriate time. Such differences in the mode of regulation may also reflect differences in the biological roles of these enzymes between vertebrate and nematode species.

4.3.4.1. Post-transcriptional regulation

Proteins which bind RNA secondary structure in the 5' and 3' UTR sequence of mature mRNA transcripts have been shown to be important for regulating the translation and stability of those transcripts. A good example of this is the iron-response element involved in regulating intracellular iron concentration. Translation of the ferritin iron-
binding protein is thought to be blocked by a regulatory protein which binds an iron-
response element in the 5′ UTR of the mature transcript of ferritin. A similar iron
response element to that found in the 5′ UTR of the ferritin mRNA is also present in the
3′ UTR of the transferrin receptor, responsible for uptake of iron in eukaryotic cells.
Binding of the same regulatory protein is thought to stabilise the mature mRNA of the
transferrin receptor, by blocking degradation of this mRNA (Casey et al., 1988).

The mature mRNA transcript of cpr-3 (1,329 bases, not including a poly(A) tail)
is the largest generated of the four genes, cpr-3, cpr-4, cpr-5 and cpr-6. Since, cpr-3
does not encode the largest protein product, this increase in size is a result of a larger 3′
UTR in the cpr-3 transcript. The FOLDRNA program generates the most stable stem-
loop structure with this 3′ UTR (-20.1 kcal at 37°C). The presence of a stable stem-loop
structure alone cannot be considered indicative of its functional role. However, the
presence of such a structure makes post-transcriptional regulation of cpr-3 expression
possible. In support of this, two cathepsin B-encoding cDNAs, hCBF1 and hCBF2,
have been isolated from a normal human embryonic fibroblast library. These cDNA
clones show differences in their 3′ UTRs from the published sequences of the human
hepatoma and kidney cathepsin B-encoding cDNAs. Both hCBF1 and hCBF2 possess a
10 bp insertion in the 3′ UTR that may allow the formation of a highly stable stem-loop
structure, absent in mRNA species without this insertion (Tam et al., 1994). The
authors indicate that the introduction of this 10 bp insertion generates a predicted stem-
loop structure with a potential free energy of -17.2 kcal, by comparison to the -8.8 kcal
stem-loop structure of cDNAs which do not contain this insertion. The authors suggest
that the increased stability of the secondary structure of the 3′ UTR may affect the
stability of this transcript. These data provide evidence that expression of cathepsin B
may be regulated post-transcriptionally in higher eukaryotes. Therefore, it is possible
that the transcript of cpr-3 may also be regulated by an analogous system. However, it
should be noted from the discussion in Section 4.3.4, that regulation of vertebrate
cathepsin B gene expression may be very different to regulation of C.elegans cathepsin
B-like gene expression, and therefore that regulation of cathepsin B at the level of
transcript stability may be specific to vertebrates
4.3.4.2. Expression of cpr-1, cpr-3, cpr-4, cpr-5 and cpr-6 may be temporally and/or spatially restricted

Sequence analysis of the 5' flank regions of cpr-3, cpr-4, cpr-5 and cpr-6 reveals clustering of several putative cis-acting regulatory elements upstream of the region of transcription initiation, determined by 5' RACE (Section 4.2.6, Figure 4.17). For cpr-4 and cpr-6, putative TATA box (TATAWAW) promoter elements were identified in the correct position with respect to transcription initiation. For cpr-3 and cpr-5, no such potential TATA box promoter sequences were identified indicating that a TATA-less promoter must be involved in transcription of these two genes. For all four genes, between two and five classic GATA motifs (WGATAR) were identified clustered in the 5' flank regions, plus additional related variants which might also bind the GATA family of transcription factors (Ko and Engel, 1993; Merika and Orkin, 1993).

GATA factors are widely recognised as being important regulators of a variety of lineage restricted genes in vertebrates. The first GATA factor, GATA-1, was originally identified as an abundant erythroid cell-specific protein that bound the WGATAR consensus sequence found in the regulatory region of several globin and non-globin erythroid specific genes (Wall et al., 1988; Evans et al., 1988; Knezetic and Felsenfeld, 1989). After cloning of GATA-1 (Tsai et al., 1989), three further GATA factors were subsequently characterised, GATA-2, GATA-3 and GATA-4 (Orkin, 1992). All four GATA factors were found to share a high degree of amino acid identity in their two zinc-finger DNA-binding domains. Binding site enrichment studies have shown that each of these GATA factors have distinct but overlapping preferences for different GATA motifs (Ko and Engel, 1993; Merika and Orkin, 1993).

The four GATA factors all have distinct but overlapping patterns of tissue specific expression. For example, GATA-1 is expressed by cells of the erythroid lineage and eosinophils (Zon et al., 1993) while GATA-2 is expressed in neutrophils and in the developing brain (Orkin, 1992), but both are expressed in mast cells and the megakaryocytic lineages.

The cis-acting GATA motif, to which GATA proteins bind, was first identified in studies of the globin genes. Numerous other erythroid genes have subsequently been isolated which require GATA elements for correct expression. Furthermore, a response
element which directs erythroid-restricted expression has been generated by combining a GATA motif with a CACCC motif upstream of a heterologous promoter (Walters and Martin, 1992).

Functionally important GATA motifs have been found in several different environments. They have been found in both promoter elements (either upstream of, or replacing, the TATA box) and within enhancer elements (Orkin, 1992). GATA factors may well perform different functions in these different settings. For example, GATA factors may inhibit expression of some genes by interacting with their core promoters and preventing formation of the pre initiation complex by steric hindrance. The core promoters of the rat platelet factor 4 (PF4) and mouse erythropoietin genes both possess GATA motifs in place of a TATA box. Initiation of transcription from these promoters requires the TATA-binding protein of TFIID and is inhibited by GATA-1 in in vitro transcription assays (Aird et al., 1994). In addition, using mobility shift assays, the authors have shown that GATA-1 interacts with the core promoter GATA motif and prevents formation of the preinitiation complex. In other settings, the binding of GATA proteins to sequences upstream of the core promoter may enhance transcription by promoting formation of the preinitiation complex (Ravid et al., 1991).

Putative GATA regulatory motifs have been implicated in the regulation of several C.elegans genes which show spatially restricted patterns of expression including; cpr-1 (Ray and McKerrow, 1992), dpy-7 (J.Gilleard, pers. comm.), the six vitellogenin genes (Spieth et al., 1985) including vit-2 (MacMorris et al., 1992), and the gut esterase gene, ges-1 (Stroher et al., 1994). Indeed, protein factors from nuclear extracts of C.elegans embryos have been shown to bind the GATA motifs of ges-1 (Stroher et al., 1994).

Several GATA factors have also been identified in C.elegans; elt-1 (Spieth et al., 1991), elt-2 (Hawkins and McGhee, 1995), elt-3 (J.Gilleard, pers. comm.) and end-1 (J.Zhu, pers. comm.). elt-1 was originally cloned by screening a C.elegans N2 genomic DNA lambda library with degenerate oligonucleotides based on the GATA-1 sequence (Spieth et al., 1991). More recently, transformation rescue experiments and sequencing of the elt-1 coding region from three C.elegans strains with non-complementing zygotic lethal mutations which define the hyd-1 locus, have provided strong evidence that hyd-1 and elt-1 are the same gene, and therefore that elt-1 plays an important role in
specification of the hypodermis (B.D.Page, pers. comm.). Furthermore, the hyd-1 mutation can be phenocopied by transformation of *C.elegans* with constructs generating elt-1 antisense RNA (K.Steward, pers. comm.). These experiments suggest that elt-1 plays a crucial role in specification of the hypodermis during embryogenesis. elt-2 was isolated from a *C.elegans* cDNA expression library using a region of the ges-1 promoter that possesses a tandem pair of GATA motifs (Hawkins and McGhee, 1995). The elt-2 gene has subsequently been shown to exhibit gut-specific expression, using *lacZ* reporter constructs (M.Hawkins, pers. comm.). Finally, the end-1 gene was cloned using a transformation rescue approach. Sequencing of this gene revealed that (like elt-1, elt-2 and elt-3) end-1 encodes a protein with significant homology to the DNA-binding domain of the GATA family of transcription factors. end-1 is thought to be important for specification of the endoderm, since end-1 mutants exhibit endodermal defects including a complete absence of the E-lineage (J.Zhu, pers. comm.). Together, these data provide a substantial amount of evidence to suggest that the GATA family of transcription factors are important regulators of a number of lineage restricted genes in *C.elegans*.

It is not possible to determine whether or not putative DNA binding motifs are functional from sequence data alone. However the clustering of a number of classic GATA motifs (and variations of this motif that may be functionally significant) in the 5′ flanks of *cpr-3, cpr-4, cpr-5* and *cpr-6* suggest that expression of these genes may also be regulated by the GATA family of transcription factors (Section 4.2.6, Figure 4.17). Furthermore, the presence of a TATA box in the 5′ flank sequences of only *cpr-4* and *cpr-6*, suggests that the mechanism of regulation of these two genes by GATA factors may be distinct from *cpr-3* and *cpr-5*. Indeed the differences in the number, type and position of GATA motifs, as well as the presence or absence of a TATA box, in the 5′ flank regions of *cpr-3, cpr-4, cpr-5* and *cpr-6* suggest that these genes may require different combinations of GATA factors for appropriate expression.

It is interesting to note that at least one of the classic GATA motifs present in the 5′ flank of each of the four genes, *cpr-3, cpr-4, cpr-5* and *cpr-6*, is a perfect VPE2 sequence. This sequence (CTGATAA) represents a subset of GATA motifs first identified in the promoter regions of all six *vit* genes of *C.elegans* which are expressed exclusively in the intestinal cells of hermaphrodites (Spieth et al., 1985). Subsequently,
the *cpr-1* cathepsin B-like gene and the *ges-1* gene, which are both expressed exclusively in the intestinal cells of *C. elegans*, have both been shown to possess at least one copy of the VPE2 motif in their respective promoter regions (Ray and McKeever, 1992; Egan et al., 1995). Furthermore, mutation analysis of the *vit-2* (MacMorris et al., 1992) and *ges-1* (Egan et al., 1995) promoters have demonstrated that the VPE2 sequences perform important roles in the expression of these two genes. Mutational analysis of the VPE2 sequences of the *vit-2* gene has revealed that these sequences perform an important role in the activation of *vit-2* in the intestine. However, these studies have also demonstrated that the VPE2 sequences may not be absolutely required, since complete deletion of the VPE2 sequences results in very low level, but correctly regulated, expression of the *vit-2* gene (MacMorris et al., 1992). In contrast, the GATA motifs in the promoter region of the *ges-1* gene, including the VPE2 sequence, appear to perform roles as both tissue specific transcriptional enhancers and silencers (Egan et al., 1995) since these sequences are required for activation of *ges-1* expression in the intestine and silencing of *ges-1* expression in the pharynx and tail. Together, these data suggest that VPE2 sequences may perform important roles in the regulation and activation of a number of gut-specific genes in *C. elegans* but that the nature of this regulation may be dependent on the context within which the GATA motifs are found in their respective promoters. Thus, the presence of at least one VPE2 sequence in the promoter regions of *cpr-1*, *cpr-3*, *cpr-4*, *cpr-5* and *cpr-6*, in conjunction with the known gut-specific expression of *cpr-1*, suggest that the four cathepsin B-like genes reported here may also exhibit gut-specific expression which may in part be regulated by the interaction of different members of the GATA family of transcription factors.

In summary, nucleotide sequence analysis of *cpr-3*, *cpr-4*, *cpr-5* and *cpr-6*, in addition to their mature transcripts, indicates these genes all have very different gene architectures. The predicted protein products of these genes display the features expected of the cathepsin B class of cysteine protease enzyme but may have different substrate specificities and activities from one another. Phylogenetic analysis, and comparison of the predicted protein sequences, indicate that the products of these genes are highly diverged and may therefore have evolved to perform distinct biological functions. Analysis of the 5' flank sequences of the four genes reveals clustering of
several putative GATA regulatory elements. The presence of these elements suggest that these genes may be temporally and/or spatially regulated.
Table 4.1

The table shows the expected sizes of the cloned 5' RACE products for *cpr-3*, *cpr-4*, *cpr-5* and *cpr-6* as determined from the sizes of the predominant 5' RACE products visualised by ethidium bromide staining of the 1.5% agarose gel (Figure 4.2). The table also shows the actual size range of the cloned 5' RACE products for each gene, determined from their nucleotide sequences.
<table>
<thead>
<tr>
<th>Gene</th>
<th>Expected size of RACE products, determined from agarose gel electrophoresis</th>
<th>Size of 5’ RACE products, determined from DNA sequence analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>cpr-3</em></td>
<td>250 bp</td>
<td>225 bp - 278 bp</td>
</tr>
<tr>
<td><em>cpr-4</em></td>
<td>210 bp</td>
<td>207 bp - 213 bp</td>
</tr>
<tr>
<td><em>cpr-5</em></td>
<td>330 bp</td>
<td>328 bp - 343 bp</td>
</tr>
<tr>
<td><em>cpr-6</em></td>
<td>270 bp and 330 bp</td>
<td>274 bp - 287 bp</td>
</tr>
</tbody>
</table>
The table summarises the data obtained using the COMPARE and DOTPLOT programs of the GCG package. These programs were used to compare the nucleotide sequences of \textit{cpr-1}, \textit{cpr-3}, \textit{cpr-4}, \textit{cpr-5} and \textit{cpr-6} to one another. The regions of homology between genes identified using the two programs are given in the table. For each region of homology, the approximate nucleotide positions delineating the beginning and end of the region are given. The numbers indicate the nucleotide position with respect to the putative ATG initiation codon for the appropriate gene (set at +1).
<table>
<thead>
<tr>
<th>gene</th>
<th>homology region (bp)</th>
<th>gene</th>
<th>homology region (bp)</th>
</tr>
</thead>
<tbody>
<tr>
<td>cpr-4</td>
<td>+240 - +490</td>
<td>cpr-5</td>
<td>+230 - +480</td>
</tr>
<tr>
<td></td>
<td>+490 - +790</td>
<td></td>
<td>+580 - +830</td>
</tr>
<tr>
<td></td>
<td>+790 - +1040</td>
<td></td>
<td>+930 - +1280</td>
</tr>
<tr>
<td>cpr-1</td>
<td>+170 - +250</td>
<td>cpr-4</td>
<td>+120 - +190</td>
</tr>
<tr>
<td></td>
<td>+300 - +550</td>
<td></td>
<td>+240 - +540</td>
</tr>
<tr>
<td></td>
<td>+800 - +1100</td>
<td></td>
<td>+690 - +1040</td>
</tr>
<tr>
<td>cpr-1</td>
<td>+150 - +300</td>
<td>cpr-3</td>
<td>+130 - +280</td>
</tr>
<tr>
<td></td>
<td>+300 - +600</td>
<td></td>
<td>+330 - +680</td>
</tr>
<tr>
<td></td>
<td>+700 - +1100</td>
<td></td>
<td>+1180 - +1580</td>
</tr>
<tr>
<td>cpr-1</td>
<td>+150 - +300</td>
<td>cpr-5</td>
<td>+80 - +230</td>
</tr>
<tr>
<td></td>
<td>+300 - +550</td>
<td></td>
<td>+230 - +480</td>
</tr>
<tr>
<td></td>
<td>+800 - +1100</td>
<td></td>
<td>+980 - +1280</td>
</tr>
<tr>
<td>cpr-4</td>
<td>+140 - +200</td>
<td>cpr-6</td>
<td>+230 - +300</td>
</tr>
<tr>
<td></td>
<td>+240 - +540</td>
<td></td>
<td>+630 - +980</td>
</tr>
<tr>
<td></td>
<td>+790 - +1040</td>
<td></td>
<td>+1180 - +1580</td>
</tr>
<tr>
<td>cpr-1</td>
<td>+300 - +650</td>
<td>cpr-6</td>
<td>+580 - +930</td>
</tr>
<tr>
<td></td>
<td>+850 - +1120</td>
<td></td>
<td>+1180 - +1530</td>
</tr>
<tr>
<td>cpr-3</td>
<td>+280 - +580</td>
<td>cpr-4</td>
<td>+240 - +540</td>
</tr>
<tr>
<td></td>
<td>+1130 - +1480</td>
<td></td>
<td>+740 - +1040</td>
</tr>
<tr>
<td>cpr-3</td>
<td>+230 - +580</td>
<td>cpr-5</td>
<td>+180 - +480</td>
</tr>
<tr>
<td></td>
<td>+1180 - +1480</td>
<td></td>
<td>+980 - +1280</td>
</tr>
<tr>
<td>cpr-6</td>
<td>+230 - +330</td>
<td>cpr-5</td>
<td>+80 - +160</td>
</tr>
<tr>
<td></td>
<td>+630 - +880</td>
<td></td>
<td>+230 - +480</td>
</tr>
<tr>
<td></td>
<td>+1180 - +1530</td>
<td></td>
<td>+980 - +1280</td>
</tr>
<tr>
<td>cpr-6</td>
<td>+630 - +930</td>
<td>cpr-3</td>
<td>+330 - +630</td>
</tr>
<tr>
<td></td>
<td>+1280 - +1530</td>
<td></td>
<td>+1230 - +1480</td>
</tr>
</tbody>
</table>
Table 4.3

The table summarises the data obtained using the GAP program of the GCG package. The program was used to generate DNA sequence alignments of the different regions of homology between cpr-1, cpr-3, cpr-4, cpr-5 and cpr-6, identified using the COMPARE and DOTPLOT programs (Table 4.2). The table shows both the maximum percentage homology and the maximum length of uninterrupted homology observed between each region analysed.
<table>
<thead>
<tr>
<th>gene</th>
<th>region of gene</th>
<th>vs.</th>
<th>gene</th>
<th>region of gene</th>
<th>region of maximum homology</th>
<th>longest length of continuous homology</th>
</tr>
</thead>
<tbody>
<tr>
<td>cpr-3</td>
<td>+130 - +280</td>
<td></td>
<td>cpr-1</td>
<td>+150 - +300</td>
<td>83.3% (25/30)</td>
<td>8 bp</td>
</tr>
<tr>
<td></td>
<td>+330 - +680</td>
<td></td>
<td></td>
<td>+300 - +600</td>
<td>91.5% (43/47)</td>
<td>20 bp</td>
</tr>
<tr>
<td></td>
<td>+1180 - +1580</td>
<td></td>
<td></td>
<td>+700 - +1100</td>
<td>88.9% (16/18)</td>
<td>11 bp</td>
</tr>
<tr>
<td>cpr-6</td>
<td>+230 - +300</td>
<td></td>
<td>cpr-4</td>
<td>+140 - +200</td>
<td>81.8% (9/11)</td>
<td>4 bp</td>
</tr>
<tr>
<td></td>
<td>+630 - +980</td>
<td></td>
<td></td>
<td>+240 - +540</td>
<td>95.7% (22/23)</td>
<td>18 bp</td>
</tr>
<tr>
<td></td>
<td>+1180 - +1580</td>
<td></td>
<td></td>
<td>+790 - +1040</td>
<td>87.9% (29/33)</td>
<td>11 bp</td>
</tr>
<tr>
<td>cpr-1</td>
<td>+150 - +300</td>
<td></td>
<td>cpr-5</td>
<td>+80 - +230</td>
<td>85.7% (18/21)</td>
<td>11 bp</td>
</tr>
<tr>
<td></td>
<td>+300 - +550</td>
<td></td>
<td></td>
<td>+230 - +480</td>
<td>89.5% (17/19)</td>
<td>12 bp</td>
</tr>
<tr>
<td></td>
<td>+800 - +1100</td>
<td></td>
<td></td>
<td>+980 - +1280</td>
<td>91.9% (34/37)</td>
<td>16 bp</td>
</tr>
<tr>
<td>cpr-4</td>
<td>+120 - +190</td>
<td></td>
<td>cpr-6</td>
<td>+240 - +490</td>
<td>93.8% (30/32)</td>
<td>17 bp</td>
</tr>
<tr>
<td></td>
<td>+300 - +550</td>
<td></td>
<td></td>
<td>+240 - +540</td>
<td>87.5% (35/40)</td>
<td>18 bp</td>
</tr>
<tr>
<td></td>
<td>+800 - +1100</td>
<td></td>
<td></td>
<td>+690 - +1040</td>
<td>100% (29/29)</td>
<td>29 bp</td>
</tr>
<tr>
<td>cpr-3</td>
<td>+230 - +480</td>
<td></td>
<td>cpr-4</td>
<td>+120 - +190</td>
<td>86.7% (13/15)</td>
<td>6 bp</td>
</tr>
<tr>
<td></td>
<td>+580 - +830</td>
<td></td>
<td></td>
<td>+240 - +540</td>
<td>87.5% (35/40)</td>
<td>18 bp</td>
</tr>
<tr>
<td></td>
<td>+930 - +1280</td>
<td></td>
<td></td>
<td>+790 - +1040</td>
<td>96.8% (30/31)</td>
<td>25 bp</td>
</tr>
<tr>
<td>cpr-3</td>
<td>+280 - +580</td>
<td></td>
<td>cpr-4</td>
<td>+240 - +540</td>
<td>94.7% (18/19)</td>
<td>14 bp</td>
</tr>
<tr>
<td></td>
<td>+1130 - +1480</td>
<td></td>
<td></td>
<td>+740 - +1040</td>
<td>80.8% (21/19)</td>
<td>11 bp</td>
</tr>
<tr>
<td>cpr-3</td>
<td>+230 - +580</td>
<td></td>
<td>cpr-5</td>
<td>+180 - +480</td>
<td>91.3% (21/23)</td>
<td>11 bp</td>
</tr>
<tr>
<td></td>
<td>+1180 - +1480</td>
<td></td>
<td></td>
<td>+980 - +1280</td>
<td>87.0% (20/23)</td>
<td>11 bp</td>
</tr>
<tr>
<td>cpr-6</td>
<td>+230 - +330</td>
<td></td>
<td>cpr-5</td>
<td>+80 - +160</td>
<td>100% (17/17)</td>
<td>17 bp</td>
</tr>
<tr>
<td></td>
<td>+630 - +880</td>
<td></td>
<td></td>
<td>+230 - +480</td>
<td>93.8% (15/16)</td>
<td>11 bp</td>
</tr>
<tr>
<td></td>
<td>+1180 - +1530</td>
<td></td>
<td></td>
<td>+980 - +1280</td>
<td>90.6% (29/32)</td>
<td>18 bp</td>
</tr>
<tr>
<td>cpr-6</td>
<td>+580 - +930</td>
<td></td>
<td>cpr-1</td>
<td>+300 - +650</td>
<td>100% (27/27)</td>
<td>27 bp</td>
</tr>
<tr>
<td></td>
<td>+1180 - +1530</td>
<td></td>
<td></td>
<td>+850 - +1120</td>
<td>87.5% (28/32)</td>
<td>12 bp</td>
</tr>
<tr>
<td>cpr-6</td>
<td>+630 - +930</td>
<td></td>
<td>cpr-3</td>
<td>+330 - +630</td>
<td>92.6% (25/27)</td>
<td>20 bp</td>
</tr>
<tr>
<td></td>
<td>+1280 - +1530</td>
<td></td>
<td></td>
<td>+1230 - +1480</td>
<td>84.2% (16/19)</td>
<td>8 bp</td>
</tr>
</tbody>
</table>
Table 4.4

The table shows the percentage similarity (SIM) and percentage identity (ID) between the prepro-enzyme and predicted mature enzyme forms of the cathepsin B-like gene products of *cpr-1, cpr-3, cpr-4, cpr-5* and *cpr-6* and of human cathepsin B. The data were obtained from alignments of the amino acid sequence of each enzyme pair using the GAP program of the GCG package.
<table>
<thead>
<tr>
<th></th>
<th>cpr-4</th>
<th>cpr-5</th>
<th>cpr-6</th>
<th>cpr-1</th>
<th>human cathepsin B</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SIM</td>
<td>ID</td>
<td>SIM</td>
<td>ID</td>
<td>SIM</td>
</tr>
<tr>
<td>cpr-3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mature</td>
<td>68.9</td>
<td>54.5</td>
<td>69.0</td>
<td>56.9</td>
<td>68.6</td>
</tr>
<tr>
<td>Pre-Pro</td>
<td>64.2</td>
<td>47.8</td>
<td>65.2</td>
<td>50.9</td>
<td>63.6</td>
</tr>
<tr>
<td>cpr-4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mature</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>83.5</td>
</tr>
<tr>
<td>Pre-Pro</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>76.9</td>
</tr>
<tr>
<td>cpr-5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mature</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>71.0</td>
</tr>
<tr>
<td>Pre-Pro</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>67.1</td>
</tr>
<tr>
<td>cpr-6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mature</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>70.5</td>
</tr>
<tr>
<td>Pre-Pro</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>63.7</td>
</tr>
<tr>
<td>cpr-1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mature</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>69.8</td>
</tr>
<tr>
<td>Pre-Pro</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>62.3</td>
</tr>
</tbody>
</table>
Figure 4.1

A). Summary of the 5' RACE methodology

1. Anneal pooled gene specific primers (GSP1) for each of the four genes to total RNA from a mixed stage population of *C. elegans*.

2. Synthesise 1st strand cDNA using Superscript II reverse transcriptase.

3. Degrade RNA using RNase H.

4. Generate dC-tailed 1st strand cDNA using dCTP and terminal deoxynucleotidyl transferase.

5. **1st Round PCR:** PCR amplify the dC-tailed 1st strand cDNA using pooled gene specific primers (GSP2) for each of the four genes and the Anchor primer.

6. **2nd Round PCR:** PCR amplify the products from the 1st Round PCR reaction using the Universal primer and one gene specific primer (GSP3) in four independent PCR reactions (1 reaction for each gene).

B). A summary of the 5' RACE protocol
A

1

2

3

4

5

Anchor Primer

6

Universal Primer

B

Mixed stage C. elegans Bristol N2 total RNA

cDNA Pool A: 1st strand cDNA synthesis using pooled GSP1 primers for the 4 genes followed by dC-tailing

1st round PCR amplification using the Anchor Primer and pooled GSP2 Primers

2nd round PCR amplification using the Anchor Primer and one GSP2 Primer for each gene in 4 independent PCR reactions

Agarose gel electrophoresis, gel purification and cloning of predominant RACE products

cDNA Pool B: 1st strand cDNA synthesis using pooled GSP1 primers for the 4 genes followed by dC-tailing

1st round PCR amplification using the Anchor Primer and pooled GSP2 Primers

2nd round PCR amplification using the Anchor Primer and one GSP2 Primer for each gene in 4 independent PCR reactions.

Agarose gel electrophoresis, gel purification and cloning of predominant RACE products
Figure 4.2

Results obtained from the 5’ RACE experiments

A. Results obtained from PCR amplification of cDNA pool A (see Figure 4.1B).
B. Results obtained from PCR amplification of cDNA pool B (see Figure 4.1B).

Lane 1: PCR products generated by the 1st round PCR amplification step

Lanes 2-5: PCR products obtained after the 2nd Round PCR amplification step

Lane 2: *cpr*-3
Lane 3: *cpr*-6
Lane 4: *cpr*-4
Lane 5: *cpr*-5

C. GibcoBRL 5’ RACE controls

Lane X: Results obtained after the 1st round PCR amplification step using the Anchor Primer, the GibcoBRL control GSP2 primer and dC-tailed cDNA template generated from 1 pg of GibcoBRL control RNA (an *in-vitro* transcribed 891bp region of the Chloramphenicol Acetyl Transerase gene engineered to include a poly(A) Tail). A 2nd round PCR amplification step was not performed with the GibcoBRL control.

Expected size of product from GibcoBRL control RNA: 738bp
Figure 4.3

The results obtained after repeating the 2nd round PCR amplification step for *cpr-6*

The 2nd round PCR amplification step was performed using the 1st round PCR amplification products from cDNA pool A (see Figure 4.1 and 4.2), the Universal Primer and the *cpr-6* GSP3 primer.
DNA Sequence of the cloned 5' RACE products of cpr-3

The figure shows the DNA sequence of each of the 5' RACE clones of cpr-3 analysed, including a 50bp region of overlap with the cpr-3 cDNA clone (cm12b6). The nucleotide sequence of the 50bp region of the cm12b6 cDNA clone to which the 5' RACE clones were compared and the nucleotide sequence of the largest 5' RACE clone sequenced is given. Thereafter, only sequence mismatches are shown, with identical bases being marked by a dash (-). The consensus sequence generated from these data is shown below the 5' RACE clone sequences. The putative ATG translation initiation codon is boxed (ATG). A second ATG codon, in-frame with the putative ATG initiation codon, which was identified in one of the eight sequenced 5' RACE clones is marked in bold (ATG). A translation termination codon in-frame with and upstream of the putative ATG translation initiation codon is underlined (TGA)
Consensus sequence

AACATCAAAATGTCGGGATCTATGGCTTAXCTCTACCCACGTTTTACCTCCGCATGTGATCTACAAGAACTGCGATC

Products of cpr-3

Sequence from the cloned 5' RACE

Sequence from cpr-3 DNA clone (cm126)

AACATCAAAATGTCGGGATCTATGGCTTAXCTCTACCCACGTTTTACCTCCGCATGTGATCTACAAGAACTGCGATC

AACATCAAAATGTCGGGATCTATGGCTTAXCTCTACCCACGTTTTACCTCCGCATGTGATCTACAAGAACTGCGATC
DNA Sequence of the cloned 5' RACE products of cpr-4

The figure shows the DNA sequence of each of the 5' RACE clones of cpr-4 analysed, including a 50bp region of overlap with the cpr-4 cDNA clone (cm14e3). The nucleotide sequence of the 50bp region of the cm14e3 cDNA clone to which the 5' RACE clones were compared and the nucleotide sequence of the largest 5' RACE clone sequenced is given. Thereafter, only sequence mismatches are shown, with identical bases being marked by a dash (-). The consensus sequence generated from these data is shown below the 5' RACE clone sequences. The putative ATG translation initiation codon is boxed (ATG)
Consensus sequence

Product from cp-4
Clone 5 RACE
Sequence from the
DNA clone (cm1+3)
Sequence from cp-1
DNA Sequence of the cloned 5’ RACE products of *cpr*-5

The figure shows the DNA sequence of each of the 5’ RACE clones of *cpr*-5 analysed, including a 50bp region of overlap with the *cpr*-5 cDNA clone (cm04d10). The nucleotide sequence of the 50bp region of the cm04d10 cDNA clone to which the 5’ RACE clones were compared and the nucleotide sequence of the largest 5’ RACE clone sequenced is given. Thereafter, only sequence mismatches are shown, with identical bases being marked by a dash (-). The consensus sequence generated from these data is shown below the appropriate region of the 5’ RACE clone sequences. The putative ATG translation initiation codon is boxed (ATG).
DNA Sequence of the cloned 5' RACE products of *cpr-6*

The figure shows the DNA sequence of each of the 5' RACE clones of *cpr-6* analysed, including a 50bp region of overlap with the *cpr-6* cDNA clone (cm01a5). The nucleotide sequence of the 50bp region of the cm01a5 cDNA clone to which the 5' RACE clones were compared and the nucleotide sequence of the largest 5' RACE clone sequenced is given. Thereafter, only sequence mismatches are shown, with identical bases being marked by a dash (-). The consensus sequence generated from these data is shown below the 5' RACE clone sequences. The putative ATG translation initiation codon is boxed (ATG).
A comparison of the predicted amino acid sequences of cpr-3, cpr-4, cpr-5 and cpr-6, together with the previously isolated cpr-1 cathepsin B-like gene and human cathepsin B (humcb)

The alignment was generated using the PILEUP program of the GCG package and some alignment by eye. The region boxed at the N-terminus indicates the hydrophobic residues of the putative pre-domain.
Figure 4.9

A comparison of the 5' flank sequence of \(cpr-3\), \(cpr-4\), \(cpr-5\) and \(cpr-6\) to the consensus DNA sequence generated by sequencing the 5' RACE clones for each of the four genes. The DNA sequences of the 5' RACE clones were compared to the DNA sequences of the 5' flank of the appropriate gene, up to the putative ATG translation initiation codon.
For each gene:

Translation Initiation Codon

\[\text{Sequence 1} \quad 5' \quad \text{Sequence 2} \quad 3'\]

\[\text{Sequence 3} \quad 5' \quad \text{Sequence 4} \quad 3'\]

\[\text{Sequence 5} \quad 5' \quad \text{Sequence 6} \quad 3'\]
Figure 4.10

Potential RNA secondary structure for the 3' UTR of the mature transcript of *cpr-3*

The structure was calculated using the FOLDRNA program of the GCG package. The free energy for the secondary structure was calculated in kilocalories (kcal) at 37°C by the FOLDRNA program.
Energy: -20.1

```
1180 GA AAGAGAGA
  1190 C GGA
1200 C A
1210 GAAUUAAACGGGAA
1220 CA AU
1230 AGU A
1240 A GA
1250 GUA
1260 A
1270 A
1280 C
1290 AA
1300 AA
1310 AGU U
1320 AU
```

UGAA
The annotated DNA sequence of cpr-3

The coding DNA sequence is indicated in capitals (CAPITALS). The amino acid residue encoded by each codon is shown below the first nucleotide of each codon. The catalytic amino acid residues, C, H and N, are marked by grey boxes (□). Non-coding sequences (introns and 5’ and 3’ flank sequences) are indicated in lowercase (lowercase). The 5’ end of each sequenced 5’ RACE clone is indicated by a vertical arrow (↑). The number above each arrow indicates the number of sequenced 5’ RACE clones whose 5’ ends terminate at that point. The double vertical lines (||) indicate the start of the poly(A) tail and the boxed nucleotides (■■■■) indicate the putative cleavage/poly(A) signal.
Figure 4.12

The annotated DNA sequence of *cpr-4*

The coding DNA sequence is indicated in capitals (*CAPITALS*). The amino acid residue encoded by each codon is shown below the first nucleotide of each codon. The catalytic amino acid residues, C, H and N, are marked by grey boxes (■). Non-coding sequences (introns and 5' and 3' flank sequences) are indicated in lowercase (*lowercase*). The 5' end of each sequenced 5' RACE clone is indicated by a vertical arrow (†). The number above each arrow indicates the number of sequenced 5' RACE clones whose 5' ends terminate at that point. The double vertical lines (∥) indicate the start of the poly(A) tail and the boxed nucleotides (■■■■) indicate the putative cleavage/poly(A) signal.
The annotated DNA sequence of *cpr-5*

The coding DNA sequence is indicated in capitals (*CAPITALS*). The amino acid residue encoded by each codon is shown below the first nucleotide of each codon. The catalytic amino acid residues, C, H and N, are marked by grey boxes (■). Non-coding sequences (introns and 5′ and 3′ flank sequences) are indicated in lowercase (*lowercase*). The 5′ end of each sequenced 5′ RACE clone is indicated by a vertical arrow (▼). The number above each arrow indicates the number of sequenced 5′ RACE clones whose 5′ ends terminate at that point. The double vertical lines (║) indicate the start of the poly(A) tail and the boxed nucleotides (■■■■) indicate the putative cleavage/poly(A) signal. The non-coding nucleotide residues marked in bold (*bold*) indicate the region of the 5′ flank of *cpr-5* which was only sequenced on one strand.
The annotated DNA sequence of *cpr-6*

The coding DNA sequence is indicated in capitals (**CAPITALS**). The amino acid residue encoded by each codon is shown below the first nucleotide of each codon. The catalytic amino acid residues, C, H and N, are marked by grey boxes (■). Non-coding sequences (introns and 5' and 3' flank sequences) are indicated in lowercase (**lowercase**). The 5' end of each sequenced 5' RACE clone is indicated by a vertical arrow (↑). The number above each arrow indicates the number of sequenced 5' RACE clones whose 5' ends terminate at that point. The double vertical lines (||) indicate the start of the poly(A) tail and the boxed nucleotides (■■) indicate the putative cleavage/poly(A) signal.
Figure 4.15

Schematic representations of the gene architectures of \textit{cpr-3}, \textit{cpr-4}, \textit{cpr-5} and \textit{cpr-6}

The diagrams are all drawn to the same scale and the scale is indicated.
Tandemly repeated hexamer sequences within the 5' flank of cpr-5

The figure shows the entire extent of the sequenced region of the 5' flank of cpr-5 and a portion of the coding region of the gene. Non-coding DNA sequence is indicated in lowercase (lowercase) and coding sequence in capitals (CAPITALS). The non-coding nucleotide residues marked in bold (bold) indicate the region of the 5' flank of cpr-5 which was only sequenced on one strand. Single units of the TTATGC hexamer repeat are boxed ([ttaggc]) and single units of the TTAGGC hexamer repeat are indicated by ovals ([ttaggc]). Hexamer repeats with single base mismatches to these two repeat sequences are also indicated. These repeats are marked by boxes or ovals, depending on which repeat they most closely resemble. The single base mismatches are underlined (underlined) and the equivalent residue of the appropriate repeat indicated above the mismatch.
Figure 4.17

Putative regulatory elements in the 5' flanks of *cpr-3, cpr-4, cpr-5* and *cpr-6*

The figure shows the DNA sequence of the 5' flanks of *cpr-3, cpr-4, cpr-5* and *cpr-6* up to the putative ATG translation initiation codon (ATG). The 5' end of each sequenced 5' RACE clone is indicated by an arrow (↑). The number above each arrow indicates the number of sequenced 5' RACE clones whose 5' ends terminate at that point. Putative regulatory sequence elements are marked in bold (bold). Putative TATA promoter elements (*TATAWAW*) are boxed and underlined ( *tatawaw* ). Classic GATA sequence motifs (*WGATAR*) are indicated by a box with one pointed end, illustrating the orientation of the motif ( *wgatar* ). Sequence elements with variations of the classic GATA motif which may bind the GATA family of transcription factors are indicated by an arrow, illustrating their orientation ( *ngatnn* ).
Figure 4.18

The phylogenetic relationships of the putative cathepsin B-like enzymes encoded by cpr-3, cpr-4, cpr-5 and cpr-6 to cysteine protease enzymes in the Genbank and EMBL databases recreated by the NEIGHBOR program

The unrooted phylogenetic tree was generated using the NEIGHBOR program of the PHYLIP PACKAGE version 3.5c (Felsenstein, 1993) and the distances were calculated according to the Dayhoff PAM matrix. The branch lengths represent the percentage of non-synonymous amino acid changes. Node A represents the archetype of all cathepsin B and B-like enzymes analyzed. Node B represents the archetype of all other cysteine protease enzymes analyzed.

KEY: Enzyme nomenclature: cp, cysteine protease; cpb, cathepsin B or B-like cysteine protease; csp, cathepsin S cysteine protease; cpl, cathepsin L cysteine protease; cph, cathepsin H cysteine protease.

<table>
<thead>
<tr>
<th>Name</th>
<th>Accession Number</th>
<th>Name</th>
<th>Accession Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>hum.cpb</td>
<td>Homo sapiens (M14221)</td>
<td>trcu.cp</td>
<td>Trypanosoma cruzi (X54414)</td>
</tr>
<tr>
<td>moused.cpb</td>
<td>Mus musculus (M14222)</td>
<td>tocng1.cp</td>
<td>Trypanosoma congolense (L25130)</td>
</tr>
<tr>
<td>chicken.cpb</td>
<td>Gallus gallus (U18083)</td>
<td>tocng2.cp</td>
<td>Trypanosoma congolense (Z25813)</td>
</tr>
<tr>
<td>bovine.cpb</td>
<td>Bos taurus (L06075)</td>
<td>tumu.cp</td>
<td>Theileria annulata (M886659)</td>
</tr>
<tr>
<td>sm31.cpb</td>
<td>Schistosoma mansoni (M21309)</td>
<td>tpav1.cp</td>
<td>Theileria parva (M67476)</td>
</tr>
<tr>
<td>sjpan.cpb</td>
<td>Schistosoma japonicum (K0708)</td>
<td>tpav2.cp</td>
<td>Theileria parva (M37791)</td>
</tr>
<tr>
<td>speri.cpb</td>
<td>Sarcophaga peregrina (D16823)</td>
<td>pfia.cp</td>
<td>Plasmodium falciparum (P25805)</td>
</tr>
<tr>
<td>lmez.cpb</td>
<td>Lethimania mexicana (A85599)</td>
<td>pfiv.cp</td>
<td>Plasmodium vivax (L26362)</td>
</tr>
<tr>
<td>taeni.cpb</td>
<td>Trichinella aestivum (X76013)</td>
<td>pviv.cp</td>
<td>Plasmodium vivax (L08050)</td>
</tr>
<tr>
<td>trusi.cpb</td>
<td>Nicotiana rustica (X83895)</td>
<td>yest.cp</td>
<td>Saccharomyces cerevisiae (M97910)</td>
</tr>
<tr>
<td>cecp1.cpb</td>
<td>Caenorhabditis elegans, cpr-1 (M74797)</td>
<td>ehist1.cp</td>
<td>Entamoeba histolytica (M94162)</td>
</tr>
<tr>
<td>cecp3.cpb</td>
<td>Caenorhabditis elegans, cpr-3 (L38980)</td>
<td>ehist2.cp</td>
<td>Entamoeba histolytica (M94163)</td>
</tr>
<tr>
<td>cecp4.cpb</td>
<td>Caenorhabditis elegans, cpr-4 (L38985)</td>
<td>ccarrcp.cp</td>
<td>Ctripras carp (L30110)</td>
</tr>
<tr>
<td>cecp5.cpb</td>
<td>Caenorhabditis elegans, cpr-5 (L38986)</td>
<td>harmer1.cp</td>
<td>Homarus americanus (X63567)</td>
</tr>
<tr>
<td>cecp6.cpb</td>
<td>Caenorhabditis elegans, cpr-6 (L38984)</td>
<td>harmer2.cp</td>
<td>Homarus americanus (X63568)</td>
</tr>
<tr>
<td>HcAc1.cpb</td>
<td>Homonchus contortus, AC1 (M31112)</td>
<td>harmer3.cp</td>
<td>Homarus americanus (X63569)</td>
</tr>
<tr>
<td>HcAc2.cpb</td>
<td>Homonchus contortus, AC2 (M60212, M60213, M34859 and M34860)</td>
<td>hum.ph</td>
<td>Homo sapiens (P09668)</td>
</tr>
<tr>
<td>HcAc3.cpb</td>
<td>Homonchus contortus, AC3 (M83388)</td>
<td>hum.pl</td>
<td>Homo sapiens (P07711)</td>
</tr>
<tr>
<td>HcAc4.cpb</td>
<td>Homonchus contortus, AC4 (M83386)</td>
<td>hum.cp</td>
<td>Homo sapiens (P25774)</td>
</tr>
<tr>
<td>HcAc5.cpb</td>
<td>Homonchus contortus, AC5 (M83385)</td>
<td>mouse.cp</td>
<td>Mus musculus (J02583)</td>
</tr>
<tr>
<td>OosterCP1.cpb</td>
<td>Ostertagia ostertagi, CP1 (M88503 and M88504)</td>
<td>agly.cp</td>
<td>Almus glutinosus (U13940)</td>
</tr>
<tr>
<td>ddisc.cpb</td>
<td>Dictyostelium discoideum (L36204)</td>
<td>achid.cp</td>
<td>Actinidia chinensis (P09785)</td>
</tr>
<tr>
<td>tthe.cpb</td>
<td>Tetrahymena thermophila (L03212)</td>
<td>bali.cp</td>
<td>barley Hordeum spp. (S45163)</td>
</tr>
<tr>
<td>lmez.cpb</td>
<td>Leishmania mexicana (X62163)</td>
<td>curi.cp</td>
<td>Cicer arietinum (X82011)</td>
</tr>
<tr>
<td>lpf.cpb</td>
<td>Leishmania pifanoi (M97695)</td>
<td>gmax1.cp</td>
<td>Glycine max (Z32795)</td>
</tr>
<tr>
<td>thru.cpb</td>
<td>Trypanosoma brucei (X16465)</td>
<td>gmax2.cp</td>
<td>Glycine max (D28876)</td>
</tr>
</tbody>
</table>

218
Figure 4.19

The phylogenetic relationships of the putative cathepsin B-like enzymes encoded by *cpr-3, cpr-4, cpr-5* and *cpr-6* to cysteine protease enzymes in the Genbank and EMBL databases, recreated by the PROTPARS program

Two equally parsimonious unrooted phylogenetic trees was generated using the PROTPARS program of the PHYLIP PACKAGE version 3.5c (Felsenstein, 1993). These two trees were identical with the exception of one cluster. Therefore, only one of the trees, and the region of the second tree that differed, is shown.

**KEY:** Enzyme nomenclature: cp, cysteine protease; cpb, cathepsin B or B-like cysteine protease; cps, cathepsin S cysteine protease; cpl, cathepsin L cysteine protease; cpb, cathepsin H cysteine protease.

<table>
<thead>
<tr>
<th>Name</th>
<th>Accession Number</th>
<th>Name</th>
<th>Accession Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>hum.cpb</td>
<td><em>Homo sapiens</em> (M14221)</td>
<td>toru.cp</td>
<td>Trypanosoma cruzi (X54414)</td>
</tr>
<tr>
<td>mousecpb</td>
<td><em>Mus musculus</em> (M14222)</td>
<td>toong1.cp</td>
<td>Trypanosoma congolense (L25130)</td>
</tr>
<tr>
<td>chik.cpb</td>
<td>Gallsus galls (U18083)</td>
<td>toong2.cp</td>
<td>Trypanosoma congolense (Z28513)</td>
</tr>
<tr>
<td>bovine.cpb</td>
<td><em>Bos taurus</em> (L06075)</td>
<td>tans.cp</td>
<td>Theileria annulata (M86659)</td>
</tr>
<tr>
<td>sm31.cpb</td>
<td>Schistosoma mansoni (M21309)</td>
<td>tpav1.cp</td>
<td>Theileria parva (M67476)</td>
</tr>
<tr>
<td>sjapon.cpb</td>
<td>Schistosoma japonicum (Y70968)</td>
<td>tpav2.cp</td>
<td>Theileria parva (M37791)</td>
</tr>
<tr>
<td>sperg.cpb</td>
<td>Sarcophaga peregrina (D16823)</td>
<td>pfar.cp</td>
<td>Plasmodium falciparum (P25805)</td>
</tr>
<tr>
<td>lmx.cpb</td>
<td>Leishmania mexicana (K48599)</td>
<td>pvi.cp</td>
<td>Plasmodium vivax (L26362)</td>
</tr>
<tr>
<td>taesti.cpb</td>
<td>Triticum aestivum (X66013)</td>
<td>pvim.cp</td>
<td>Plasmodium vinckeii (L08500)</td>
</tr>
<tr>
<td>arru.cpb</td>
<td>Nicotiana rustica (X81995)</td>
<td>yeast.cp</td>
<td>Saccharomyces cerevisiae (M97910)</td>
</tr>
<tr>
<td>ccpp1.cpb</td>
<td>Caenorhabditis elegans, cpr-1 (M74797)</td>
<td>ehist1.cp</td>
<td>Entamoeba histolytica (M94162)</td>
</tr>
<tr>
<td>ccpp3.cpb</td>
<td>Caenorhabditis elegans, cpr-3 (L39890)</td>
<td>ehist2.cp</td>
<td>Entamoeba histolytica (M94163)</td>
</tr>
<tr>
<td>ccpp4.cpb</td>
<td>Caenorhabditis elegans, cpr-4 (L39895)</td>
<td>ccorp.cp</td>
<td>Cyprinus carpio (L30111)</td>
</tr>
<tr>
<td>ccpp5.cpb</td>
<td>Caenorhabditis elegans, cpr-5 (L39896)</td>
<td>hamer1.cp</td>
<td>Homarus americanus (X63567)</td>
</tr>
<tr>
<td>ccpp6.cpb</td>
<td>Caenorhabditis elegans, cpr-6 (L39894)</td>
<td>hamer2.cp</td>
<td>Homarus americanus (X63568)</td>
</tr>
<tr>
<td>HoA1.cpb</td>
<td><em>Haemonchus contortus</em>, AC1 (M31112)</td>
<td>hamer3.cp</td>
<td>Homarus americanus (X63569)</td>
</tr>
<tr>
<td>HoA2.cpb</td>
<td><em>Haemonchus contortus</em>, AC2 (M60212, M60213, M34859 and M34860)</td>
<td>hum.cpb</td>
<td>Homo sapiens (P09668)</td>
</tr>
<tr>
<td>HoA3.cpb</td>
<td><em>Haemonchus contortus</em>, AC3 (M80388)</td>
<td>hum.cpl</td>
<td>Homo sapiens (P097711)</td>
</tr>
<tr>
<td>HoA4.cpb</td>
<td><em>Haemonchus contortus</em>, AC4 (M80386)</td>
<td>hum.cps</td>
<td>Homo sapiens (P25774)</td>
</tr>
<tr>
<td>HoA5.cpb</td>
<td><em>Haemonchus contortus</em>, AC5 (M80385)</td>
<td>mouse.cp</td>
<td>Mus musculus (J02583)</td>
</tr>
<tr>
<td>OosterCP1.cpb</td>
<td>Ostertagia ostertgi, CP1 (M88503 and M88504)</td>
<td>aghy.cp</td>
<td>Almus glutinosus (U13940)</td>
</tr>
<tr>
<td>ddisc.cpb</td>
<td>Dictyostelium discoide (L36204)</td>
<td>aghd.cp</td>
<td>Actinidia chinensis (P00785)</td>
</tr>
<tr>
<td>the.cpb</td>
<td>Tetrahymena thermophila (L03212)</td>
<td>boral.cp</td>
<td>barley, Hordeum spp. (S45163)</td>
</tr>
<tr>
<td>lmx.cpb</td>
<td>Leishmania mexicana (X62163)</td>
<td>cari.cp</td>
<td>Cicer arietinum (X82011)</td>
</tr>
<tr>
<td>lplc.cpb</td>
<td>Leishmania pifanoi (m97695)</td>
<td>gmao1.cp</td>
<td>Glycine max (Z32795)</td>
</tr>
<tr>
<td>thur.cpb</td>
<td><em>Trypanosoma brucei</em> (X16465)</td>
<td>gmao2.cp</td>
<td>Glycine max (D28576)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>osat1.cp</td>
<td>Oryza sativa (P25778)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>osat2.cp</td>
<td>Oryza sativa (X80876)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>vst1.cp</td>
<td>Vicia sativa (Z30338)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>vst2.cp</td>
<td>Vicia sativa (Z34895)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>vst3.cp</td>
<td>Vicia sativa (X75749)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>vst4.cp</td>
<td>Vicia sativa (Z34899)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>znays1.cp</td>
<td>Zea mays (D45403)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>znays2.cp</td>
<td>Zea mays (D45403)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ntata1.cp</td>
<td>Nicotiana tabacum (Z13964)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ntata2.cp</td>
<td>Nicotiana tabacum (Z13965)</td>
</tr>
</tbody>
</table>
trees generated by the PROTPARS program

Region of variation of the two equally parsimonious

C. elegans putative cathespin B-like enzymes
The phylogenetic relationships of the putative cathepsin B-like enzymes encoded by cpr-3, cpr-4, cpr-5 and cpr-6 to cathepsin B and B-like enzymes in the Genbank and EMBL databases

A and B). Two different representations of the same unrooted majority-rule consensus phylogenetic tree generated using the PROTPARS program of the PHYLIP package version 3.5c (Felsenstein, 1993). Bootstrap resampling was performed with the PROTPARS program and the majority-rule consensus tree was obtained from the 20 trees generated by the bootstrap analysis using the CONSENSUS program of the PHYLIP package, version 3.5c. In Figure 4.20B, the number by each grouping represents the percentage of the 20 trees generated by Bootstrap resampling that support that particular grouping.

C). Unrooted phylogenetic tree generated using the NEIGHBOR program of the PHYLIP PACKAGE version 3.5c. The distances were calculated according to the Dayhoff PAM matrix. The branch lengths represent the percentage of non-synonymous amino acid changes.

KEY: Enzyme nomenclature: cpb, cathepsin B or B-like cysteine protease

The grey ellipses highlight the three groups of putative C.elegans cathepsin B-like enzymes generated by the NEIGHBOR and PROTPARS programs.

<table>
<thead>
<tr>
<th>Name</th>
<th>Accession Number</th>
<th>Name</th>
<th>Accession Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>hum.cpb</td>
<td>Homo sapiens (M14221)</td>
<td>cecpr5.cpb</td>
<td>Caenorhabditis elegans, cpr-3 (L39896)</td>
</tr>
<tr>
<td>moused.cpb</td>
<td>Mus musculus (M14222)</td>
<td>cecpr6.cpb</td>
<td>Caenorhabditis elegans, cpr-6 (L39894)</td>
</tr>
<tr>
<td>chicken.cpb</td>
<td>Gallus gallus (U18083)</td>
<td>HcAC1.cpb</td>
<td>Haemonchus contortus, AC1 (M31112)</td>
</tr>
<tr>
<td>bovine.cpb</td>
<td>Bos taurus (L06075)</td>
<td>HcAC2.cpb</td>
<td>Haemonchus contortus, AC2 (M60212, M60213, M34859 and M34860)</td>
</tr>
<tr>
<td>sm31.cpb</td>
<td>Schistosoma mansoni (M21309)</td>
<td>HcAC3.cpb</td>
<td>Haemonchus contortus, AC3 (M80388)</td>
</tr>
<tr>
<td>sjapon.cpb</td>
<td>Schistosoma japonicum (X70968)</td>
<td>HcAC4.cpb</td>
<td>Haemonchus contortus, AC4 (M80386)</td>
</tr>
<tr>
<td>spereg.cpb</td>
<td>Sarcoptes peregrinus (D16823)</td>
<td>HcAC5.cpb</td>
<td>Haemonchus contortus, AC5 (M80385)</td>
</tr>
<tr>
<td>lmez.cpb</td>
<td>Leishmania mexicana (x48599)</td>
<td>OosterCP1.cpb</td>
<td>Ostertagia ostertagi, CP1 (M88503 and M88504)</td>
</tr>
<tr>
<td>taesti.cpb</td>
<td>Triticum aestivum (X66013)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>nrusti.cpb</td>
<td>Nicotiana rustica (X81995)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>cecpr1.cpb</td>
<td>Caenorhabditis elegans, cpr-1 (M74797)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>cecpr3.cpb</td>
<td>Caenorhabditis elegans, cpr-3 (L39890)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>cecpr4.cpb</td>
<td>Caenorhabditis elegans, cpr-4 (L39895)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
An alignment of the predicted amino acid sequences of cpr-1, cpr-3, cpr-4, cpr-5, cpr-6, AC-2, CP-1 and mouse cathepsin B

The alignment was generated using the PILEUP program of the GCG package. Vertical bars (|) mark intron positions with respect to the amino acid sequence. cpr-1, cpr-3, cpr-4, cpr-5 and cpr-6 encode the cathepsin B-like genes of C.elegans. AC-2 encodes a cathepsin B-like gene from H.contortus. CP-1 encodes a cathepsin B-like gene from O.ostertagi.

A-F denote intron positions conserved between genes that are not only conserved with respect to the amino acid sequence but also to the exact nucleotides at which intron / exon boundaries occur within codons. These positions are also marked by an arrow (▼). For AC-2 and CP-1, all 11 intron positions show this level of conservation.

H denotes a site of possible intron loss

I-K denote sites where introns are found in close proximity to the intron sites defined by AC-2 and CP-1 but are not at identical positions

L and M denote sites where introns are found in close proximity to one another but outwith the intron sites defined by AC-2 and CP-1
Chapter 5
Chapter 5
The Temporal and Spatial Expression Patterns of
$cpr-3$, $cpr-4$, $cpr-5$ and $cpr-6$

5.1. Introduction

In order to further characterise $cpr-3$, $cpr-4$, $cpr-5$ and $cpr-6$, I decided to analyse their temporal and spatial patterns of expression. The temporal and spatial expression pattern of a gene can provide a substantial amount of information regarding its biological function. For example, expression restricted to certain cell types (or tissues) may indicate a specific function for that gene in those cells or tissues. In such circumstances, the biological roles associated with a particular cell or tissue type may help assign a potential biological role for the gene. By analogy, it may also be possible to infer the biological role of a gene from its temporal expression pattern. For example, a gene which is expressed within a defined period of time during embryogenesis may be required for embryogenesis. When such expression patterns are matched with the DNA sequence of the gene analysed, it may be possible to accurately predict the biological function of that gene. For example, DNA sequencing might reveal the isolation of a collagen gene, while expression analysis might reveal that the gene is expressed cyclically in the hypodermis prior to each larval moult. Together, such information would provide strong evidence of a structural role for this gene in the *C. elegans* cuticle, which could not be obtained from the DNA sequence or expression data alone.

When considering gene families, the level of complexity is much increased since the biological functions performed by the gene family may require the interaction of none, some or all members of the gene family to varying degrees. Thus, at one extreme, all members of a gene family may be required to act in concert in order to perform their biological roles while, at the other extreme, all members of a gene family may act independently of one another to perform distinct biological roles. Together, determining the DNA sequence of the members comprising a gene family as well as determining their spatial and temporal expression patterns may help to determine the degree of interaction between the various genes comprising the gene family. The results obtained so far for
the cathepsin B-like gene family of *C. elegans* provide a good example for illustrating this. From the results discussed in Chapters 3 and 4, there is a substantial amount of evidence to suggest that the cathepsin B-like genes of *C. elegans* are highly diverged and encode enzymes with different substrate specificities and activities. Accordingly, the discovery that these genes are expressed in the same tissue but at different stages during *C. elegans* development would suggest a requirement for different enzymes, or combinations of enzymes, to process or degrade different substrates, or combinations of substrates, at different stages during *C. elegans* development. Alternatively, the discovery that all these genes are expressed in the same tissue and at the same time would suggest that the enzymes may be required at the same time to effectively broaden the spectrum of substrates which can be digested or processed at any one time. In both these examples, different members of the gene family may have to interact substantially in order to perform their functions. For example, a particular substrate might require the activity of several different isoenzymes for its complete digestion or processing. In other scenarios, the genes may be expressed in different tissues, either at the same time or at different times. Such situations may reflect a requirement for slightly different enzymes in different cellular environments which raises the possibility that these genes may have different biological roles by virtue of their involvement in distinct cellular processes.

None of the hypothetical situations, including those discussed above, are mutually exclusive. Accordingly, large gene families may exhibit complex temporal and spatial expression patterns, with different members of a gene family interacting with each other to different extents. Comparison of different expression patterns may allow the members comprising a gene family to be ordered into subgroups on the basis of the frequency with which members of the same group are expressed at the same time in the same tissues, or on the basis of which members are never expressed together. Such analysis may be extremely useful for understanding how members of a gene family interact. For example, there is now evidence to suggest that, though cuticular collagen genes are expressed in the hypodermis, different groups of these genes are expressed at slightly different times prior to moultling (I.L. Johnstone, pers. comm.). It is thought that those collagen genes expressed simultaneously are capable of producing functional heterotrmeric proteins required for cuticle formation, but are not capable of doing so with collagens expressed
at different stages. This may therefore represent a mechanism by which incompatible members of a gene family are effectively isolated from one another.

To conclude, I decided to analyse the temporal and spatial expression patterns of *cpr*-3, *cpr*-4, *cpr*-5 and *cpr*-6. Such studies, in conjunction with the DNA sequence data for the four genes, would provide information essential for understanding both the biological roles and biological interactions of the protein products encoded by these genes.

5.2. Analysing the temporal expression patterns of *cpr*-3, *cpr*-4, *cpr*-5 and *cpr*-6

5.2.1. Experimental approaches used

I decided to use Northern blot analysis and semi-quantitative reverse transcription PCR (s-q rtPCR), to study the temporal expression patterns of *cpr*-3, *cpr*-4, *cpr*-5 and *cpr*-6. Northern blot analysis is one of the simplest and most widely used approaches for assaying the temporal expression patterns of genes during development. This approach is known to be reliable and standard protocols for this method are readily available for *C. elegans*. Developmental Northern blot analysis requires the availability of cloned genes that are constitutively expressed during the life cycle. These clones can be used to control for equal loading in different lanes of RNA isolated from different developmental stages, allowing comparisons of transcript abundance to be made between samples. Such cloned genes are available for use as controls for both total RNA and poly(A)$^+$ selected RNA extracted from *C. elegans*. The pCe7 clone, which contains a single repeat of the 18s and 26s ribosomal DNA sequences (Files and Hirsh, 1981), and the pCeIF clone, which contains a cDNA encoding the *C. elegans* homologue of the eukaryotic initiation factor 4A (Roussell and Bennett, 1992), are recommended for use as loading controls for total RNA and poly(A)$^+$ selected RNA, respectively.

Northern blot analysis is relatively insensitive when compared to other approaches. The lack of sensitivity necessitates generating large synchronous populations of *C. elegans* in order to obtain sufficient material for analysis from each developmental stage. Though the generation of synchronous populations of *C. elegans* can be achieved relatively easily by the isolation of eggs (Sulston and Hodgkin, 1988)
and utilisation of the L₁ arrest caused by starving, the degree of synchrony decreases dramatically with each subsequent generation. Thus, large numbers of eggs must be obtained initially in order to generate a sufficiently large synchronous population such that all the desired RNA samples can be extracted within one generation. Since the culture of extremely large populations of *C. elegans* is unfeasible in the laboratory, there is an upper limit on the total number of eggs that can be obtained in order to generate a synchronous population. Furthermore, for reasons which remain unclear, we have found that better synchrony is obtained with small populations rather than with large, dense cultures. These factors limit the maximum size of the synchronous population which can be generated. In addition to the requirement for large amounts of material, the insensitivity of Northern blot analysis also sets a limit on the frequency with which a synchronous population can be sampled, since sampling too frequently will result in insufficient material for analysis. A high frequency of sampling is essential if temporal patterns of expression of a target gene are to be accurately assayed. If sampling becomes too infrequent, or the sample window too large, transcripts from genes with very restricted temporal patterns of expression may be missed, giving rise to inaccurate results. For these reasons I decided to use Northern blot analysis to assay the transcript abundance of *cpr-3, cpr-4, cpr-5* and *cpr-6* from only two samples, embryonic RNA and mixed stage RNA. This analysis would give some insight into the temporal pattern of expression of each of these genes. More importantly, it would reveal both the sizes of the transcripts generated and whether or not each of these genes generates multiple mRNA size species.

Semi-quantitative reverse transcriptase polymerase chain reaction (s-q rtPCR) was also used to assay the temporal expression patterns of the four genes. This approach incorporates a PCR amplification step which makes it much more sensitive than Northern blot analysis, however the technique is also inherently more complex.

In brief, the method I used involves converting RNA from each sample to cDNA, using reverse transcriptase. Two sets of PCR primers; one set, specific to the target transcript and the other, to a constitutively expressed transcript acting as an internal control, are then used to amplify the cDNAs simultaneously in a single reaction. The number of PCR cycles is chosen empirically to ensure that reactants are still in excess after amplification, and do not therefore become rate limiting. After amplification, the
PCR products are electrophoresed through an agarose gel and Southern blotted to a nylon filter. The filter is probed using sequences specific to the target and control genes. The resulting autoradiograph is subsequently used as a template to excise the regions of the filter corresponding to the amplification products of the cDNA molecules derived from the target and control transcripts. The amount of radioactivity from each band is then assayed by counting in scintillant. The abundance of the target gene can therefore be expressed as a ratio of the signal obtained from the target gene to that of the control gene. This method therefore measures the relative abundance of a target transcript with respect to a previously chosen constitutively expressed reference transcript as an internal control. The method gives no indication of the absolute abundance of any given transcript. However, it can be used to plot fluctuations of that transcript's abundance in different samples of RNA, for example a time course, by comparing the test gene to internal control gene relative abundance ratio between samples.

The PCR amplification step responsible for the much increased sensitivity of s-q rtPCR is also a potential source of error, since many factors can affect the efficiency of amplification when using PCR, especially small variations in the relative concentrations of the reactants within each tube. This problem is largely overcome by introducing an internal control transcript for the PCR reaction, since most factors which affect the efficiency of amplification of cDNA derived from the target transcript will also affect the efficiency of amplification of cDNA derived from the control transcript. Accordingly, s-q rtPCR requires a control gene which is constitutively expressed, to act both as an internal control for the PCR reaction and to allow comparisons to be made between RNA samples from different stages of C.elegans development. The ama-1 gene, which encodes the large subunit of RNA polymerase II, was chosen as such a control. The mRNA accumulation of this gene is believed to be relatively constant during the larval and adult stages of development (D.Riddle, pers. comm.), and s-q rtPCR experiments performed in our laboratory with a number of cuticular collagen genes, including dpy-7, sqt-1 and col-12, using ama-1 as a control support this assumption (I.L.Johnstone, pers. comm.).

The amplification step of s-q rtPCR makes this method much more sensitive than Northern blot analysis. Therefore, sufficient RNA for analysis can be isolated at much more frequent intervals from small, highly synchronous populations of C.elegans.
ability to sample at a higher frequency than is feasible for Northern blot analysis reduces the possibility of overlooking highly restricted temporal patterns of expression of a test gene. For example, the L1 stage lasts approximately 11 hours if worms are incubated at 25°C. During this time the worms are continuously developing, thus isolation of a single RNA sample covering the L1 phase from a highly synchronous population will not be representative of all stages of L1 development because different stages will be enriched depending on the time at which the RNA was extracted after hatch. Indeed, experience in our laboratory using Northern blot analysis to study the expression of cuticular collagen genes has demonstrated that expression of these genes can be missed when using wide sample windows in conjunction with highly synchronous cultures of *C. elegans*.

This problem can be reduced dramatically by increasing the frequency with which RNA samples are taken, so that each sample represents a much narrower window of time during development. For example, RNA extracted at two hour intervals after hatch would generate at least 5 samples covering the L1 phase of development. Together, these samples will be more representative of the period of *C. elegans* development which occurs during this phase and will increase the probability of detecting transcripts from genes with highly restricted temporal patterns of expression. Furthermore, even if such RNA samples are pooled to generate samples representing each of the larval and adult phases of development, each pooled sample will be more representative of the developmental phase it covers than a single RNA sample isolated during the same phase, because there will be less enrichment for any one period of development within that phase. These factors make s-q rtPCR a powerful tool for the analysis of temporal patterns of gene expression. The approach was therefore chosen because of the reasons discussed above and because the approach had already been used very successfully in our laboratory to analyse the temporal patterns of expression of a number of cuticular collagen genes.

5.2.2. Results of Northern blot analysis

Total RNA was prepared from embryos (the embryos were not staged and therefore represent mixed stage embryos) or mixed stage N2 *C. elegans* populations
(containing RNA from embryos, larvae and adults). Four aliquots of 4μg of mixed stage or embryo total RNA were electrophoresed through a 1% agarose gel under denaturing conditions and blotted to Amersham Hybond N. The filter was subsequently cut into four strips, each carrying a sample of mixed staged and embryo total RNA. The Northern blots were hybridised overnight with the cDNA clones of cpr-3, cpr-4, cpr-5 and cpr-6. The filters were initially washed at medium stringency and subsequently at high stringency. The autoradiographic results obtained after medium stringency and high stringency washes are shown in Figures 5.1 and 5.2 respectively. Previous results, from Southern analysis with C.elegans genomic DNA, using the same cDNA clones as probes (Chapter 3, Section 3.2.5), demonstrated that wash stringencies similar to the high stringency wash of the Northern blots were sufficient to prevent significant cross hybridisation of cpr-3, cpr-4, cpr-5 and cpr-6 to one another or to other sequences in the C.elegans genome. Thus, the results obtained from the Northern blots after high stringency washing are unlikely to be a result of cross-hybridisation to additional, related sequences. The hybridisation pattern obtained after medium stringency washing (Figure 5.1) is virtually identical to that obtained after high stringency washing (Figure 5.2), the only difference being the much reduced signal intensity obtained after high stringency washing. This was probably a result of a general decrease in the amount of probe hybridising to target transcript caused by the high stringency washes (and not as a result of reduced-cross hybridisation to other, similar transcripts from related genes) since the actual pattern of hybridising bands remains unchanged. Two features of the hybridisation patterns obtained support this conclusion. First, though transcripts from cpr-4, cpr-5 and cpr-6 were detected after high stringency washing, no transcript was detected for cpr-3 with medium stringency washing, suggesting that medium stringency washing is at least sufficient to prevent cross hybridisation of a cpr-3 probe to transcripts from the other 3 genes. Second, though the cpr-6 transcript is noticeably larger than both the cpr-4 and cpr-5 transcripts, none of the Northern blots produce two hybridising bands, indicating that cross-hybridisation of the cpr-6 probe to cpr-4 and cpr-5 transcripts (or vice versa) is not occurring. It seems extremely unlikely that a cDNA probe, for example the cpr-6 cDNA, would cross hybridise to another transcript of very similar size to that of cpr-6 without also cross-hybridising to the transcripts of either cpr-3, cpr-4 or cpr-5. Furthermore, the results obtained from Southern blot analysis of C.elegans genomic
DNA using the four cDNA clones as probes (Chapter 3, Section 3.2.2) did not reveal any indication of the presence of additional cathepsin B-like genes with extensive homology to \textit{cpr-3}, \textit{cpr-4}, \textit{cpr-5} or \textit{cpr-6} within the genome of \textit{C.elegans}. Accordingly, I will principally discuss the results obtained after medium stringency washing.

Ethidium bromide staining of the RNA gel (Figure 5.3A) prior to transfer to the Hybond N filter indicated approximately equal loading of mixed stage (lane 1) and embryo (lane 2) total RNA. Transfer of the RNA to the Nylon filter was also tested by ethidium bromide staining of the agarose gel after blotting, which revealed no visible fluorescence. The loading and transfer of RNA was also confirmed for two of the four filters by probing with a 7 kb \textit{Bam}HI rDNA fragment containing a single repeat of the 18s and 26s ribosomal DNA sequences, gel purified from the pCe7 clone (Files and Hirsh, 1981). The results from this (Figure 5.3B) indicate that there was slightly less mixed stage RNA than embryo RNA present on the filters and that both the mixed stage and embryo RNA was degraded. However, it also demonstrated that the total RNA had been successfully transferred to the Nylon filter in these two cases, and therefore that the transfer of RNA to the other two strips was probably also successful.

Single hybridising bands, corresponding to a transcript size of approximately 1.3 kb for \textit{cpr-4} and \textit{cpr-5} and of approximately 1.45 kb for \textit{cpr-6} are visible on the autoradiographs (Figures 5.1B, 5.1C and 5.1D), indicating that these three genes each generate predominant transcripts of a single size. This is consistent with the results obtained from the 5' RACE experiments (Chapter 4, Section 4.2.6), which did not generate any evidence for transcription initiation occurring at more than one promoter region for each gene. The sizes of the mature transcripts of \textit{cpr-4}, \textit{cpr-5} and \textit{cpr-6} (predicted from DNA sequencing of their cDNA and 5' RACE clones) are 1,120 bases, 1,152 bases and 1,260 bases, respectively. As expected, these sizes are consistently smaller than the sizes obtained from Northern blot analysis since poly(A) tails were not included in the mature transcript sizes predicted by DNA sequencing. Furthermore, Northern blot analysis indicates that the mature transcript of \textit{cpr-6} is approximately 150 bases larger than the mature transcripts of \textit{cpr-4} and \textit{cpr-5}, which is also in agreement with the predicted mature transcript sizes of these three genes determined by sequencing their cDNA and 5' RACE clones.
Hybridising bands corresponding to the transcripts of cpr-4 and cpr-5 were only visible with mixed stage total RNA (Figures 5.1B and 5.1C, lane 1). This suggests that induction of expression of these two genes occurs at some point around or after hatching of the embryo. A hybridising band corresponding to the cpr-6 transcript is visible in both embryo and mixed stage RNA at medium stringency (Figure 5.1D). However, the signal intensity of the transcript in mixed stage RNA is much greater than that in the embryo RNA, suggesting that the abundance of the cpr-6 transcript is much greater in mixed stage RNA than in embryo RNA. It should be noted that, by the nature of the population of C.elegans from which the mixed stage RNA was isolated, some embryo RNA was also present. However, since the cpr-6 transcript abundance seen with embryo RNA is so low, the contribution made by embryo RNA to the signal seen in the mixed stage RNA must be minimal. Therefore, the increased cpr-6 transcript abundance seen with the mixed stage RNA must be due to an increase in expression occurring at some time around or after hatch.

Since no transcript was detectable for cpr-3 from total RNA (Figure 5.1A), poly(A)$^+$ RNA was selected from embryo and mixed stage total RNA. There was insufficient poly(A)$^+$ selected RNA to calculate the concentration by spectrophotometry and therefore all the poly(A)$^+$ RNA obtained was electrophoresed through a 1.0% agarose gel and transferred to an Amersham Hybond N membrane. The poly(A)$^+$ Northern blot was hybridised overnight with a riboprobe generated from the cloned cpr-3 cDNA (cm12b6). The Northern blot was then washed at high stringency prior to exposure to medical X-ray film. The results of this are shown in Figure 5.4. The blot was subsequently stripped and hybridised with a riboprobe generated from pCelIF to control for equal loading. This clone contains a cDNA insert encoding the C.elegans homologue of the eukaryotic initiation factor 4A (Roussell and Bennett, 1992) and is thought to represent one of the better controls for equal loading of poly(A)$^+$ RNA. Unfortunately, I was unable to detect any signal with this and therefore could not control for equal loading. Consequently, it was not possible to make comparisons of cpr-3 transcript abundance between embryo and mixed stage samples.

The Northern blot analysis (Figure 5.4) reveals that cpr-3 has a single predominant species of mature transcript. This is consistent with the 5' RACE experiments performed with cpr-3, which suggest that transcription of cpr-3 is initiated
from a single promoter region upstream of this gene. The size of the mature transcript is approximately 1.6 kb. This is larger than the transcript size of 1,329 bases determined by sequencing the cDNA and 5' RACE clones of cpr-3 and is explained by the presence of a poly(A) tail in the mature transcript which was not included in the transcript size determined by sequencing.

In summary, Northern blot analysis revealed that the predominant transcripts of cpr-3, cpr-4, cpr-5 and cpr-6, are each of a single size species. For each gene, the size of the mature transcript and the presence of a single size species of mature transcript is in agreement with data obtained from the 5' RACE experiments and sequencing of the appropriate 5' RACE and cDNA clones for each gene. For cpr-4, cpr-5 and cpr-6, the Northern blot data also suggest that induction of expression of these genes occurs around, or after, the time of embryo hatch, since transcript abundance is greatly increased in the mixed stage total RNA sample. Such data could not be determined for cpr-3 because controls for equal loading failed to generate hybridising signals.

5.2.3. Results from sq-rtPCR analysis

5.2.3.1. Initial analysis of the temporal expression patterns of cpr-3, cpr-4, cpr-5 and cpr-6 using sq-rtPCR

Initially, the s-q rtPCR approach outlined in Section 5.2.1 was used to analyse the relative transcript abundance of cpr-3, cpr-4, cpr-5 and cpr-6 in the 20 staged cDNA samples generated by Dr I.L.Johnstone (see Chapter 2, Section 2.13). These samples were generated from total RNA isolated at two hour intervals during the four larval and early adult stages of C.elegans development. For each gene, 23 cycles of PCR were performed using cDNA template from each of the 20 samples and primer pairs which produced recognisable size differences between the PCR products generated from cDNA and contaminating genomic DNA. This size difference was achieved by selecting primers specific to coding sequence either side of introns. The nucleotide sequences and the positions of each primer pair are given in Chapter 2, Section 2.13.1. The entire PCR reaction for each time point was electrophoresed through an agarose gel, Southern blotted to Hybond N (Amersham International plc) and probed using the same primers as
used for the PCR amplification step (5' end-labelled using polynucleotide kinase). The filters were washed at medium stringency prior to exposure to medical X-ray film. The levels of radioactivity were not assayed by scintillation and accordingly all steps outlined in Section 5.2.1 subsequent to the generation of each autoradiograph were omitted.

Only 23 cycles of PCR were used to ensure that the PCR reactions were stopped whilst still in the log phase of amplification. This would reduce the probability of misrepresentation of relative abundance caused by premature depletion of primer pairs in the PCR reactions. For each gene, 23 cycles of PCR yielded very low amounts of each product, with the products being either not or barely visible on ethidium bromide stained agarose gels prior to Southern blotting. In contrast, 35 cycles of PCR using a mix of all 20 cDNA samples and the primer pair for cpr-3 generates clearly visible bands (data not shown). This suggests that the amplification obtained after 23 cycles of PCR is sufficiently low such that the reactants are still in excess post amplification and are therefore not rate limiting.

One concern with the amplification step of s-q rtPCR is that excessive amplification of inappropriate template may affect the quantitative nature of this approach, by acting as a sink for the PCR primers and so reducing their effective concentration in the reaction mixture. Therefore, the same oligonucleotides used for the PCR amplification steps were also used as probes for hybridisation to the Southern blots of the PCR amplified products. This approach allows all the more abundant PCR amplification products, whether derived from appropriate or inappropriate template, to be visualised. The results obtained from this Southern blot analysis (Figures 5.5, 5.6, 5.7 and 5.8) reveal that all the predominant hybridising bands are of the expected size for amplification from the cDNA and genomic DNA templates of the appropriate target gene and of ama-1. This suggests that each primer pair is capable of specifically amplifying the appropriate template. For the same reason, these data suggest that none of the target gene primer pairs are incompatible with the ama-1 primer pair, since none of these combinations behaved specifically in combined reactions (i.e. they did not produce additional predominant products other than those expected). However, the cpr-4 primer pair did produce several weakly hybridising bands of inappropriate sizes. Though none of these bands individually generated strong hybridising signals, the number of bands
produced suggested that amplification of inappropriate templates could potentially act as a sink for the cpr-4 primer pair.

The autoradiographic data (Figures 5.5, 5.6, 5.7 and 5.8) reveal several important features regarding the relative transcript abundance of cpr-3, cpr-4, cpr-5 and cpr-6 during larval and adult stages. By comparing the signal intensity of the band corresponding to the transcript of the target gene to that of the ama-1 internal control, the relative transcript abundance of the target gene in each sample can be assessed. Such comparison at each time point for each gene reveals that there is no dramatic increase in expression occurring for any of the four genes cpr-3, cpr-4, cpr-5 and cpr-6, between any two adjacent time points during the larval and adult stages of *C. elegans* development. In contrast, the autoradiographic data obtained for the collagen gene *dpy*-7, using the same approach (Figure 5.9, from I.L. Johnstone), reveals a large increase in *dpy*-7 transcript levels, with respect to *ama-1*, which recurs at a similar time point prior to molting at each larval stage. Therefore it appears that none of the four genes, cpr-3, cpr-4, cpr-5 or cpr-6, show a rapid induction of expression at any point within the series, nor do they show any evidence of cyclical patterns of expression during *C. elegans* development. Interestingly, the signal intensity of the band corresponding to the cpr-3 transcript shows less variation through the larval and adult stages of *C. elegans* development than *ama-1*. This suggests that cpr-3 transcript abundance may fluctuate less during *C. elegans* larval and adult development than *ama-1*.

In summary, the initial s-q rtPCR experiments indicate that the parameters chosen are appropriate for analysis of the temporal expression patterns of cpr-3, cpr-4, cpr-5 and cpr-6. The results generate no evidence of highly restricted temporal expression patterns for the four genes, with no rapid and short-lived increases in relative transcript abundance occurring within any one developmental stage. The results also show no evidence of a cyclical pattern of expression during the larval and adult stages of *C. elegans* development. Such patterns have been observed for several cuticular collagen genes, including *dpy*-7 (Figure 5.9) which imply induction at one point within each larval stage. Since cuticular collagens are required for synthesis of the new cuticle prior to each larval moult, the periodic expression patterns observed with those cuticular collagen genes analysed, suggest that each larval moult may follow a similar programmed sequence of events. The cathepsin B enzyme has been shown to be capable of degrading
collagen and is thought to perform a role in bone remodelling in vertebrates (Chapter 1, Section 1.7.4). Accordingly, it may have been possible that the cathepsin B-like gene family evolved in nematode species to perform an analogous role in remodelling the cuticle. However, the absence of such a cyclical pattern of expression for cpr-3, cpr-4, cpr-5 and cpr-6 during the larval stages of development suggest that this is not the case.

5.2.3.2. Adaptations made to the initial s-q rtPCR protocol

From the initial s-q rtPCR experiments, it is apparent that cpr-3, cpr-4, cpr-5 and cpr-6 do not show the type of expression pattern observed with dpy-7. With none of the genes showing rapid induction of expression, highly temporally restricted expression within any one developmental stage or cyclical expression. However, it is possible that these genes might possess more subtle temporal expression patterns during C. elegans development, with gradual changes in the relative transcript abundance occurring over several samples. Accordingly, the s-q rtPCR protocol was adapted to enable experiments to be performed which could determine if such gradual fluctuations were occurring.

In order to analyse subtle changes in transcript abundance between stages, it is essential that real differences in transcript abundance can be distinguished from differences arising from experimental error. This is especially important with s-q rtPCR because the introduction of a PCR amplification step makes this protocol inherently more complex than approaches such as Northern blot analysis, thereby increasing the probability of introducing error. Accordingly, the s-q rtPCR protocol was adapted to address this problem and the adaptations made are discussed below.

The s-q rtPCR approach was simplified to make the procedure more reliable. Accordingly, the number of cDNA samples analysed in each experiment was reduced. The initial s-q rtPCR experiments indicate that none of the four genes exhibited highly restricted patterns of temporal expression within any one developmental stage. Therefore, the cDNA samples covering each stage of development were pooled to generate five amalgamated samples each representing the L1, L2, L3, L4 or adult stage, respectively. An equal volume of cDNA from each time point was amalgamated as follows: samples 1-6 for L1, samples 7-9 for L2, samples 10-12 for L3, samples 13-15 for
L4 and samples 16-20 for early adult. Equal volumes of each appropriate cDNA sample were used to generate the amalgamated samples because the results from the initial s-q rtPCR experiments (Figures 5.5, 5.6, 5.7 and 5.8) suggest that no one sample contains significantly higher concentrations of cDNA than any other. First, the fluctuations in signal intensity corresponding to ama-1 cDNA template between different samples are not consistent among the four genes, suggesting that such fluctuations do not reflect differences in the cDNA concentration in each sample, but rather differences in the efficiency of amplification. Second, the signal intensities corresponding to cpr-3 cDNA template remain remarkably constant between the 20 samples (Figure 5.5), suggesting that each of the cDNA samples from different time points contain similar amounts of cpr-3 template. Such similar levels of cpr-3 template in each sample suggest that the transcript abundance of this gene remains relatively constant during the larval and adult stages of C. elegans development. Accordingly, this suggests that each sample contains similar quantities of cDNA.

A new primer pair was designed for cpr-4 (Chapter 2, Section 2.13.1), because the previous primer pair generated some non-specific amplification products. Since cpr-4 possesses only a single, small (49 bp) intron, the small size difference between the amplification products derived from cDNA template and contaminating genomic DNA would make it difficult to separate these products by agarose gel electrophoresis. Accordingly, the primer pair was designed to only amplify cDNA template, by designing one of the primers to span the intron site within the cDNA. The new cpr-4 primer pair was tested for its ability to specifically amplify cpr-4 cDNA template in the presence of ama-1 primers using 35 cycles of PCR, mixed stage cDNA and the same annealing temperature to be used in the s-q rtPCR experiments. The PCR reaction generated three PCR products. The sizes of these products were as expected for amplification of the cpr-4 cDNA template by the cpr-4 primer pair and amplification of the ama-1 cDNA template and contaminating genomic DNA by the ama-1 primer pair, indicating that this new primer pair was capable of specifically amplifying cpr-4 cDNA in the presence of ama-1 primers.

The hybridisation protocol of s-q rtPCR was altered to allow higher stringency washes because previous strategies using 5' end-labelled oligonucleotide probes produced filters with background signals that were too high and too variable across the
surface of the filter to allow accurate analysis by scintillation counting. Therefore, probes for \textit{cpr-3, cpr-4, cpr-5, cpr-6} and \textit{ama-1} were generated by PCR amplification (Chapter 2, Section 2.13.3) using the same primer pairs to be used in the s-q rtPCR experiments. The larger size of these probes allowed the use of high stringency washes which reduced the levels and variability of background radioactivity after hybridisation of the probes to the Southern blots of the PCR amplified products.

A control for efficient transfer of the PCR amplified DNA to the Hybond N (Amersham International plc) filters during Southern blotting was introduced to ensure that variation in the relative transcript abundance between samples was not a result in variation in the efficiency of transfer of the PCR products to the filters. Therefore, the size of each PCR reaction was doubled (to a 50\,\mu\text{l} final volume). After the amplification step, the reaction from each time point was split into two 20\,\mu\text{l} aliquots to generate two sets of samples from the same PCR reaction. Each set of samples was electrophoresed through an agarose gel and transferred to Hybond-N (Amersham International plc) by Southern blotting. The filters were then probed and washed in identical conditions prior to exposure to medical X-ray film. By comparing the autoradiographs generated by the two blots derived from the same PCR reaction, it was possible to assess the efficiency of transfer since each blot should give virtually identical results.

An additional cDNA sample covering the embryo stages of \textit{C.elegans} development was introduced because Northern blot analysis suggests that induction of expression of \textit{cpr-4, cpr-5} and \textit{cpr-6} occurs around or after the time of embryo hatch and the 20 staged cDNA samples used in the initial s-q rtPCR experiments only covered the larval and adult stages of development. First strand cDNA was generated from 2\,\mu\text{g} of mixed stage embryo RNA (Chapter 2, Section 2.13). The cDNA was then diluted to an appropriate concentration for use in s-q rtPCR. This dilution was determined using 23 cycles of PCR with primer pairs for \textit{ama-1} and \textit{cpr-3} and 10 fold serial dilutions of the mixed stage embryo cDNA. The PCR products were electrophoresed through an agarose gel and visualised by ethidium bromide staining. This identified faint PCR products generated from the 100 fold dilution and therefore a 500 fold dilution was used in the s-q rtPCR experiments. It is important to note that there is no currently available method for synchronising \textit{C.elegans} embryos. It is therefore not possible to determine what proportion of each embryo stage is represented in the embryo cDNA sample. This
sample is therefore distinct from the amalgamated samples in this respect. However, the RNA was extracted immediately after isolation of the embryos and therefore is unlikely to contain RNA from L₁ larvae, since only a very small fraction of the embryos, at the end of embryo development, will have had time to progress to the L₁ stage.

Finally, for each gene, the s-q rtPCR experiments were performed in triplicate to assess the reproducibility of each set of s-q rtPCR experiments. This allowed the use of statistical tests to help determine whether differences in relative transcript abundance between samples reflected genuine differences or merely experimental error.

5.2.3.3. Summary of the modified s-q rtPCR approach used

There follows a brief summary of the experimental procedure used to obtain the temporal expression patterns of cpr-3, cpr-4, cpr-5 or cpr-6 (the procedure is also outlined in Figure 5.10). Three independent sets of PCR reactions (A, B and C) were performed with primers specific for the target gene and for ama-1 using the mixed stage embryo cDNA and the amalgamated cDNA samples for L₁, L₂, L₃, L₄ and adult stages. Each set of PCR reactions was split into two 20 µl aliquots and electrophoresed through an agarose gel. The agarose gel was then cut into two, with each half containing amplification products from each of the six developmental stages, derived from the same set of PCR reactions. Each half of the gel was then Southern blotted to Hybond N filters (Amersham International plc) to give a total of six filters carrying the amplification products from three independent PCR reactions. Therefore, for each gene, reaction A yields filters 1 and 2, reaction B yields filters 3 and 4 and reaction C yields filters 5 and 6. The six filters for each gene were simultaneously hybridised with probes specific to the target gene and to ama-1. The filters were washed at high stringency and exposed overnight with autoradiographic film. The resulting six autoradiographs were used as a template to excise the hybridising bands from the filters which corresponded to the target gene cDNA and ama-1 cDNA amplification products. For each sample, the radioactivity of the two bands was measured by counting in scintillant. The relative transcript abundance of the target gene was expressed as the ratio of the counts per minute (cpm) obtained from the target gene cDNA amplification products to the cpm obtained from the ama-1 cDNA amplification products.
5.2.3.4. The modified s-q rtPCR approach generates reproducible results

The autoradiographic data obtained from the six filters for each of the four genes are shown in Figures 5.11, 5.12, 5.13 and 5.14. The data obtained from scintillation counting, including the relative transcript abundance ratios, for each of the four genes are shown in Tables 5.1, 5.2, 5.3 and 5.4. The relative abundance ratios obtained are summarised graphically in Figure 5.15.

For each gene, each pair of filters, from a common PCR reaction, shows a very similar pattern of hybridisation on the autoradiographs (Figures 5.11, 5.12, 5.13 and 5.14), including the relative intensity of signals between different bands on the same filter. This similarity is also reflected in the graphical representations of the relative transcript abundance ratios (Figure 5.15) where each pair of graphs shows similar trends. This indicates that the PCR products from each reaction were transferred to the Nylon filters in a reproducible manner during Southern blotting. Therefore, the variation in transcript abundance of each gene during *C. elegans* development, discussed later, is not a result of variations in transfer of the PCR products. While each pair of autoradiographs shows a very similar pattern of hybridisation, the overall signal intensity may differ substantially. This cannot be easily explained since each pair of filters was hybridised with the same probe in a single hybridisation bottle. This phenomenon does not appear to affect the results since this variation occurs to a similar extent across the entire filter.

To assess the reproducibility of the amplification step of s-q rtPCR, three independent pools of cDNA from each developmental stage were amplified in different PCR reactions. Again, looking at the autoradiographic results (Figures 5.11, 5.12, 5.13 and 5.14), it is apparent that each filter pair shows a very similar pattern of hybridisation to the other filter pairs obtained from different PCR reactions for the same gene. The three sets of s-q rtPCR reactions were capable of generating not only a similar pattern of bands after hybridisation but also a similar pattern in the relative intensities of these bands, with respect to other bands on the same autoradiograph. This reproducibility is also reflected in the graphical representations of the relative transcript abundance ratios (Figure 5.15). Each graph contains three data sets, obtained from three independent
PCR reactions (A, B and C) for each gene. Each data set produces a similar trend in the temporal pattern of expression for a given gene and therefore indicates that the PCR amplification step is reproducible, as suggested by the autoradiographic data. The scintillation samples from filters 3 and 5 for cpr-6 were lost and therefore graphical comparison of data obtained from these filters was not possible. However, the original autoradiographic data (Figure 5.14) clearly demonstrate that the three independent PCR reactions for this gene generated very similar results, suggesting that the PCR amplification step for cpr-6 is also reproducible. These data provide strong evidence that the PCR amplification step of s-q rtPCR generates reproducible results for each gene.

5.2.3.5. The temporal expression patterns of cpr-3, cpr-4, cpr-5 and cpr-6 obtained from the modified s-q rtPCR approach

As discussed above, the controls implemented suggest that the transfer of DNA to Hybond N (Amersham International plc) with the modified s-q rtPCR approach is reliable. Therefore, one set of triplicate results (obtained from the three independent PCR reactions, A, B and C) was used to analyse the temporal expression pattern of each gene. Thus, the scintillation data obtained from filters 1, 3 and 5 were analysed for each of the three genes, cpr-3, cpr-4 or cpr-5. For cpr-6, there was no scintillation data for filters 3 and 5 and therefore filters 2, 4 and 6 were analysed. The mean and standard deviation of the relative transcript abundance ratio was determined for each developmental stage, from the three independent PCR reactions performed for each gene (Tables 5.1, 5.2, 5.3 and 5.4). These results are shown graphically in Figure 5.16 and reveal some very interesting differences between the temporal expression patterns of each gene.

The results suggest that all four genes show an induction of expression at the time of hatch but this induction varies in size between the four genes. Subsequent to this induction, the genes show quite different temporal patterns of expression. Thus, maximal expression of cpr-3 occurs at L₁, while cpr-4 shows two peaks at L₁ and L₄, cpr-5 shows one major peak at L₂ and a smaller peak at L₄, and cpr-6 shows one peak in the adult. I used Student’s t-test to determine if these peaks of expression were significant. This test uses, as its null hypothesis, the assumption that two samples have
been derived from the same population, and therefore that any differences in the mean and standard deviation of these samples can be accounted for by sampling errors. Using this test, I compared the relative transcript abundance of each stage with the next stage in development (i.e. embryo with $L_1$, $L_1$ with $L_2$, $L_2$ with $L_3$, $L_3$ with $L_4$, $L_4$ with adult). The results of this are shown in Table 5.5. The $t$-test confirms most of the observations made by eye. For $cpr$-3, the most obvious variation is the increase in transcript abundance between the embryo and $L_1$ stages, which is also found to be significant using the $t$-test (Table 5.5). None of the other fluctuations were found to be significant, supporting the initial s-q rtPCR data (Figure 5.5) which suggest that $cpr$-3 transcript abundance remains fairly constant through the larval and adult stages of *C.elegans* development. The temporal expression pattern for $cpr$-4 is very similar to that of $cpr$-3 with the exception of a marked decrease in relative transcript abundance between the $L_4$ and adult stages. These data suggest that $cpr$-4 is expressed predominately in the larval stages of *C.elegans* development. The $t$-test supports this conclusion, since it identifies the increase in abundance between the embryo and $L_1$ stages and the decrease in abundance between the $L_4$ and adult stages as being highly significant and significant respectively (Table 5.5), but does not find any other variations to be significant. For $cpr$-5, there is a dramatic (11.7 fold) increase in relative transcript abundance between the embryo and $L_1$ stages and the relative transcript abundance continues to rise to a maximum in the $L_2$ stage. There is a general downward trend (interrupted by a small peak at $L_4$) in transcript abundance for $cpr$-5 in the subsequent larval and adult stages. The $t$-test found the 11.7 fold increase in relative transcript abundance between the embryo and $L_1$ stages to be significant. However, it also found all the differences in relative transcript abundance between subsequent stages to be significant. This is surprising because there is only a maximum 2.3 fold variation in relative transcript abundance between these subsequent stages. $cpr$-6 shows a very different temporal pattern of expression to the other three genes. There is a slight increase in relative transcript abundance between the embryo and $L_1$ stages however this is overshadowed by a much greater increase between the $L_3$ and $L_4$ stages which continues to rise to a maximum in the early adult. This was confirmed using the $t$-test which did not find the increase between the embryo and $L_1$ stages as significant but identified the increase between $L_3$ and $L_4$ stages as being highly significant. The temporal expression profile for
cpr-6 therefore suggests that this gene is not expressed at significant levels until around the L₃ stage, unlike the other three genes where induction appears to occur around the time of embryo hatch.

There are not only variations in the timing of expression of cpr-3, cpr-4, cpr-5 and cpr-6 but also in the degree of induction that occurs. It should be noted, before discussing these variations, that the scales of the Y-axes of the graphs shown in Figure 5.16 vary to allow comparison of the four genes. The four genes can be divided into two groups on the basis of their degree of induction. cpr-3 and cpr-4 show only 2.3 and 5.3 fold increases between minimal and maximal expression while cpr-5 and cpr-6 show 27.0 and 17.9 fold increases, respectively. Therefore, for cpr-3, the differences in the relative transcript abundance seen during C.elegans development may only represent slight variation from a gene that is essentially constitutively expressed. In contrast, though cpr-4 does not show great variation between minimal and maximal transcript abundance, the data here suggest that this gene is not constitutively expressed, but upregulated during the larval stages of development. In contrast, cpr-5 and cpr-6 are highly developmentally regulated. cpr-5 appears to be induced around hatch, and its relative transcript abundance increases dramatically up to the L₂ stage, where it begins to decrease. cpr-6 shows quite a different pattern, with a dramatic induction of expression occurring at the late L₃ / L₄ stage, and relative transcript abundance increasing into the early adult stage.

It is not possible to quantify differences in absolute transcript abundance between genes using a s-q rtPCR approach since there are many factors which can influence the relative efficiency of amplification of the target transcript with respect to the ama-1 transcript. In particular, the efficiency with which primer pairs can amplify their target sequence has a significant effect on the perceived abundance of that sequence with respect to ama-1. However, it is interesting to note that cpr-5 and cpr-6, which both show the highest degrees of induction, also have the highest relative transcript abundance ratios with respect to ama-1. At the developmental stage where maximal transcript abundance is seen for cpr-5 and cpr-6, they show a 7.6 and 3.1 fold greater abundance over ama-1. The correlation between greater degrees of induction and levels of expression with respect to ama-1 does suggest that cpr-5 and cpr-6 transcripts may be more abundant than those of cpr-3 and cpr-4, once induction has occurred.
5.2.4 Comparison of the results obtained from Northern and s-q rtPCR analysis

It is essential that a constitutively expressed gene is used as the internal control for s-q rtPCR. Though \textit{ama-1} mRNA abundance appears to remain relatively constant during the larval and adult stages of \textit{C.elegans} development, recent reports suggest that \textit{ama-1} transcripts may be more abundant in the embryo (D.Riddle, pers. comm.). If \textit{ama-1} levels increased, or decreased, dramatically during any of the stages of \textit{C.elegans} development one would expect to see a conserved dip, or peak, in the relative transcript abundance as measured by this method occurring at the same stage for each of the four genes. No such conserved dip, or peak, in expression levels is seen for these genes during the larval and adult stages indicating that \textit{ama-1} levels remain relatively constant during these stages. However, all four genes show an increase in expression between the embryo and \textit{L}_1 stages. This could represent a real induction event occurring around the time of hatch, or an artefact caused by increased \textit{ama-1} transcript levels in the embryo, giving rise to an artefactual perceived increase at hatch. The increase in relative transcript abundance between the embryo and \textit{L}_1 stages for \textit{cpr-4} and \textit{cpr-5} identified by s-q rtPCR is confirmed by the results obtained from Northern blot analysis. Northern blot analysis for \textit{cpr-4} and \textit{cpr-5} (Figures 5.1B and 5.1C) detected no evidence of transcripts for these two genes in the embryo. In contrast, transcripts for both genes were detected in mixed stage RNA. The s-q rtPCR analysis detected transcripts for both these genes in all stages of \textit{C.elegans} post hatch. For \textit{cpr-4}, this expression was predominantly in the larval stages while, for \textit{cpr-5}, high levels of expression were seen in both the larval and adult stages. Together, the data suggest that both these genes are expressed at very low levels in embryos and at much higher levels in the \textit{L}_1 stage. This provides strong evidence that induction of expression for these genes is occurring at some point around the time of hatch.

For \textit{cpr-6}, transcripts were detected in both embryo and mixed stage RNA by Northern blot analysis (Figure 5.1D). However, the Northern blot data suggested that the transcript was in far greater abundance in the mixed stage RNA. The s-q rtPCR experiments for \textit{cpr-6} generate two peaks of expression, one small peak at \textit{L}_1 and a much larger peak starting between the \textit{L}_3 and \textit{L}_4 stages and rising in the early adult
stages. Statistical analysis using the t-test suggest that only the large peak starting between the L₃ and L₄ stages is significant. Together, the Northern blot and s-q rtPCR data suggest that cpr-6 is transcribed at a very low level through the embryo and early larval stages and then induced at some point around the L₃ stage.

For cpr-3, the Northern blot results could not be used to obtain any information regarding variation in transcript abundance between embryo and mixed stage RNA samples because the control for equal loading failed to generate a signal. The s-q rtPCR results suggest that this gene is expressed constitutively during the larval and adult stages of C.elegans development. In addition, there is only a 2.3 fold increase in relative transcript abundance between the embryo and L₁ stages. These data alone suggest that cpr-3 is essentially constitutively expressed during all stages of C.elegans development, although there may be a slight increase in expression between the embryo and L₁ stages.

In summary, the Northern blot and s-q rtPCR data indicate that each of the four genes cpr-3, cpr-4, cpr-5 and cpr-6 show different patterns of temporal expression during C.elegans development. cpr-3 is expressed at all stages of development and shows only a slight increase of expression at hatch. cpr-4 appears to be induced around hatch and is expressed predominantly in the larval stages. cpr-5 and cpr-6 both show highly developmentally regulated patterns of expression. Both of these genes show over 15 fold increases in their transcript abundance between minimal and maximal expression but the timings of these inductions are quite different. cpr-5 is induced around the time of hatch, and shows relatively high levels of expression in the subsequent larval and adult stages. cpr-6 appears to be transcribed at a low level for most stages of C.elegans development but is subsequently induced around the L₃ stage, with expression levels continuing to rise through the L₄ and early adult stages. The results reported here indicate that cpr-3, cpr-4, cpr-5 and cpr-6 encode enzymes which are expressed during different stages of C.elegans development. Two possible situations best explain the variation in the temporal expression patterns observed between the four genes. In one situation, these genes might encode cathepsin B-like enzymes which perform distinct biological processes in different tissues, during different stages of C.elegans development. In the other situation, the enzyme products may be expressed in the same tissues and may perform the same or similar biological functions, but are required in different combinations during C.elegans development. The diverged predicted amino
acid sequences of the cathepsin B-like enzymes encoded by the four genes (Chapter 4) are in agreement with either of these situations, since in both cases one would expect the enzymes to have different activities and substrate specificities, either to degrade / process different substrates in different tissues or to degrade / process different substrates (or combinations of substrates) in the same tissue but at different stages during C.elegans development.

5.3. Analysing the spatial expression patterns of cpr-3, cpr-4, cpr-5 and cpr-6

5.3.1. Overview of the lacZ fusion transgene approach

lacZ fusion transgenes were used to analyse the spatial expression patterns of cpr-3, cpr-4, cpr-5 and cpr-6. In brief, the 5' flank and part of the coding region of each of the four genes was ligated upstream of, and in-frame with, the bacterial lacZ β-galactosidase gene, to produce translational gene fusions. Plasmid constructs containing these gene fusions, along with a marker plasmid, pRF4, were microinjected into the syncytial gonad of adult hermaphrodite worms. Single F2 progeny which transmitted the marker plasmid were subsequently selected and transgenic lines generated from these. The method used (Mello et al., 1991) generates a large extrachromosomal array made up mostly of tandem repeats of the construct and the pRF4 marker plasmid. Analysis of such arrays has demonstrated that they vary in size, being made up of a total of approximately 80 - 300 copies of the injected plasmids (Stinchcomb et al., 1985). The arrays are not inherited in a Mendelian manner because they are not attached to chromosomes. However, transgenic lines can be kept indefinitely by selecting for the mutant phenotype produced by the marker plasmid.

5.3.2. Advantages of using lacZ transgene fusions

The lacZ transgene fusion approach is the method of choice for initial analysis of the spatial expression patterns of genes in C.elegans. The popularity of this method is based on several reasons. First, meaningful results are obtained with a high frequency. Second, the approach allows analysis of the regulatory regions responsible for
controlling expression of the genes being studied. Third, a large number of genes have been tested using lacZ transgene fusions resulting in this approach being well characterised. Finally, extensive use of this method has revealed that it is relatively free of serious artefacts.

For a number of genes, the spatial expression pattern generated by lacZ transgene fusions has also been compared to the spatial expression pattern of the endogenous gene by using other approaches such as in-situ hybridisation or antibody staining. In many cases the spatial expression patterns produced by the gene fusions accurately reflect the spatial expression patterns of their endogenous genes. For example, lacZ fusions with the MyoD family homologue hlf-1 faithfully reproduce the tissue-specific expression pattern of the endogenous gene, in body wall muscle precursors (Krause et al., 1994). Okkema et al (1993) have also demonstrated that lacZ gene fusions with unc-54, myo-1, myo-2 and myo-3 all mimic the tissue specific expression patterns of their endogenous counterparts. The unc-54::lacZ and myo-3::lacZ fusions exhibit expression in the body-wall muscle and specialised body muscles associated with the intestine, the vulva and the uterine sheath while the myo1::lacZ and myo-2::lacZ fusions exhibit expression only in pharynx muscle. Hamelin et al (1992) have demonstrated that a lacZ::mec-7 β-tubulin gene fusion is expressed specifically in the touch neurones of transgenic worms which also react with anti-mec-7 antibodies. In other cases, the pattern of expression of test gene::lacZ fusions have correlated well with mutant phenotypes resulting from mutations in the endogenous gene whose expression is being analysed. For example, the cuticular collagen gene dpy-7, which has a dumpy mutant phenotype resulting from cuticular defects, has been shown to be expressed in the hypodermis, the site of cuticle synthesis (I.L.Johnstone, pers.comm.). More recently, Tabish et al (1995) have demonstrated that lacZ gene fusions with the osm-3 kinesin gene exhibit expression in chemosensory neurons open to the environment. The authors have correlated this pattern of expression with the osm-3 mutant phenotype, which shows structural defects in the amphid and phasmid sensilla. Such examples, suggest that in many cases lacZ gene fusions with a gene of interest accurately reflect the pattern of expression of the endogenous gene.

Other techniques are available for analysing the temporal expression patterns of genes. The two principle methods are detection of RNA in-situ (in-situ hybridisation) and antibody staining of whole, or sectioned, C.elegans. Whole worm in-situ
hybridisation has been used to analyse the spatial expression patterns of genes within embryos. However, this approach has proved problematic when used to detect RNA from developmental stages other than embryos. This is thought to be a result of the cuticle acting as a barrier which prevents equal access of the nucleic acid probes to different tissues within the worm. Antibody staining is a very powerful method because it is capable of determining the spatial distribution of the final protein product of a gene, as opposed to determining only where a gene is expressed. In addition, immunogold labelling of antibodies and electron microscopy allow the subcellular distribution of target proteins to be determined. However, these approaches require the generation of specific immunological reagents. The generation of such reagents may be very time consuming and there is no way of guaranteeing that the antibodies obtained will be specific. Furthermore, this problem can be exacerbated when studying a multigene family as it may not be possible to generate antibodies specific for each member of the family.

5.3.3. Problems associated with the use of lacZ transgene fusions

Though lacZ reporter gene fusions often reliably reflect the expression patterns of endogenous copies of the genes being tested, a few problems with this approach have been observed. One of the most commonly observed problems is that of mosaic expression. This is in part due to the maintenance of the transgene as an extrachromosomal array in transgenic worms. Because these arrays are not attached to chromosomes, they may be lost during somatic cell divisions as a result of random segregation between daughter cells, and therefore may give rise to mosaic animals. Such mosaicism will cause variation in the cells which express the transgenes. However, variable staining patterns may also arise as a result of other factors such as incomplete fixation or the inability of all the cells which should stain to express the transgene at sufficient levels to be detected. Since such factors generally result in variable expression patterns similar to those generated by 'true' mosaicism, this term is often used to describe variable staining patterns with lacZ transgene fusions although the true basis of the variability may not be known. Mosaicism (staining variability) has been reported for a number of different gene fusions studied in C.elegans. For example, the P-
glycoprotein genes, \textit{pgp-1} and \textit{pgp-3} (Lincke \textit{et al.}, 1993) and the cuticular collagen gene \textit{dpy-7} (J. Gilleard, pers. comm.) have both been shown to exhibit mosaic expression as \textit{lacZ} gene fusions in the intestine and hypodermis respectively. It is important to note that valid conclusions regarding the tissue-specific expression pattern of an endogenous gene may still be made from data obtained using \textit{lacZ} gene fusions that exhibit mosaic expression. Such conclusions can be made because mosaicism generally results in staining of different cells from the same tissue in different animals of the same transgenic line. By comparing a number of different animals from the same line, therefore, it is possible to determine the tissues that express the fusion transgene, but not particular cells within the tissue that express the fusion transgene.

In some cases, ectopic expression of \textit{lacZ} gene fusions in additional tissues outside those normally expressing the endogenous gene has also been reported. For example, ectopic expression of \textit{hlh-1} (which is expressed in the mature body wall muscle) has been observed in the pharynx, posterior larval gut and in embryonic hypodermal precursors, when using very short promotor sequences and \textit{lacZ} transgene fusions (Krause \textit{et al.}, 1994). Such ectopic expression may result from the absence of all the regulatory sequences necessary for appropriate expression of the transgene fusion when using very short 5' flank sequences. Alternatively, it is thought that such ectopic expression may also arise from interaction between sequences in the plasmid backbone and those in the promoter region being analysed when in close proximity to one another.

Another problem to be considered when using \textit{lacZ} gene fusions is the apparent inability of such fusions to exhibit expression in the embryonic, larval or adult germ line of \textit{C. elegans}. The analysis of expression of the small heat shock (\textit{hsp16}) genes using \textit{hsp16::lacZ} gene fusions has revealed that \textit{lacZ} fusions can be expressed in virtually all tissues of the worm, including cells of the gut, nervous system hypodermis and pharynx, however in no instance has germ line expression been demonstrated (Fire \textit{et al.}, 1990; Stringham \textit{et al.}, 1992). Furthermore, other transgenes, such as \textit{msp-1} which encodes a sperm specific product, have also failed to be expressed in this nematode species.

Two other problems are associated with the use of \textit{lacZ} transgene fusions: first, no transgene expression has been demonstrated in pre-twelve-cell embryos of \textit{C. elegans} to date; second, expression in a single tissue or population of cells may not be seen with
the transgene, though the endogenous gene is known to be expressed in that region (A.Fire, pers. comm.).

The problems discussed above have lead to the very recent introduction of a new set of modified lacZ reporter vectors (A.Fire, pers. comm.). These vectors essentially differ from the old vectors by the addition of one to twelve introns into the lacZ coding region. The addition of these introns can stimulate expression by two to three orders of magnitude making these vectors substantially more sensitive than the previous set. In situ localisation experiments using unc-54::lacZ constructs containing 0,1,8 and 12 intron sequences indicates that the addition of intron sequences into the lacZ coding region results in a large increase in the amount of unc-54::lacZ transcripts seen in the cytoplasm. This suggests that introns may increase the efficiency of transport of transcripts out of the nucleus, or may increase processing efficiency and/or transcript stability. In addition to the increased sensitivity, these new vectors also exhibit reduced levels of mosaicism, or staining variability, compared to the previous set of vectors. However, these vectors have not overcome the problems of ectopic expression or lack of expression in the germ line encountered with lacZ fusion transgenes. Indeed, the increase in sensitivity has also resulted in increased background expression in the pharyngeal and gut tissues, especially with vectors containing 12 introns in the lacZ coding region. However, the use of a decoy minigene (discussed in the following section) upstream of the 5' Polylinker has eliminated such background expression from promoterless lacZ fusion vectors by preventing readthrough transcription.

5.3.4. lacZ fusion vectors

Both vectors used to construct the reporter gene fusions are members of a modular set of lacZ fusion vectors (Fire et al., 1990). These vectors are well characterised because many of the studies of gene expression in C.elegans using lacZ transgene fusions have utilised them. The basic vector consists of a pUC19 derived plasmid backbone, a 5' multiple cloning site (MCS) or polylinker and the E.coli trpC::lacZ gene fusion. A number of cassettes can then be added to this, containing one of the following; a synthetic intron, an ATG start codon and nuclear localisation signal, a synthetic transmembrane segment and finally, a 3' MCS, the unc-54 3' end cassette or
the SV40 early 3' end cassette. I used the two *lacZ* fusion vectors, pPD21.28 and pPD95.03 (see Figure 5.17). The pPD21.28 vector comprises a pUC19 plasmid backbone, a 5' polylinker, a synthetic intron and an SV40 nuclear localisation signal all upstream of the *E.coli trpS::lacZ* β-galactosidase coding region. An *unc-54* 3' end supplies the cleavage / polyadenylation signal downstream of the coding region. The pPD 95.03 vector is based on the pPD21.28 vector and therefore shares many similarities. The two important differences are the insertion of twelve synthetic introns into the *E.coli TrpS::lacZ* β-galactosidase coding region. The insertion of introns into the *lacZ* gene has resulted in an increase in sensitivity by two to three orders of magnitude, making this vector significantly more sensitive than pPD21.28. The increase in sensitivity has also resulted in an increase in background, especially in the pharynx and the gut. To combat this, a decoy sequence has also been inserted just upstream of the 5' polylinker. The decoy sequence consists of a short intron followed by a short open reading frame and a consensus *C.elegans* translational start. The open reading frame terminates just upstream of the 5' end of the polylinker. Such a decoy sequence has been shown to prevent background expression from promoterless vectors by preventing readthrough transcription originating from the plasmid backbone (A.Fire, pers. comm.). It should be noted that the presence of a nuclear localisation signal in the two *lacZ* fusion vectors used will result in localisation of the fusion protein to the nucleus. Therefore such localisation does not reflect the actual subcellular localisation of the endogenous proteins encoded by the genes being analysed.

### 5.3.5. Generation of the *lacZ* gene fusions

Several factors require consideration when generating *lacZ* gene fusions. First, a sufficient amount of 5' flanking sequence for each gene must be inserted into the *lacZ* fusion vector to ensure that all regulatory elements required for correct expression of these genes are present in the flanking sequences. Promotor analysis studies with other genes such as *dpy-7* (J.Gilleard, pers. comm.), *vit-2* (MacMorris *et al.*, 1992) and *mec-3* (Xue *et al.*, 1992) suggest that as little as 250bp - 350bp of 5' flank sequence may be sufficient for driving appropriate spatial expression of a number of *C.elegans* genes. However, I decided to use substantially larger 5' flank regions from *cpr-3, cpr-4, cpr-5*
and cpr-6 for initial analysis of the spatial expression patterns of these genes. The use of larger 5’ flank regions places the lacZ fusion transgene in a more natural environment. It increases the probability of obtaining all the necessary regulatory elements to drive correct expression of each gene. It also reduces the likelihood of ectopic expression by increasing the distance between the promoter elements in the 5’ flank and sequences in the plasmid backbone. The interaction of promoter elements and sequences in the plasmid backbone may occur when these sequences are in close proximity to one another. Such interactions are thought to cause some of the ectopic expression patterns observed when using very short 5’ flank regions (Krause et al., 1994). Thus, I decided to use as much 5’ flank sequence as possible from the cloned genes, cpr-3, cpr-4, cpr-5 and cpr-6. Accordingly, 1.4 kb, 2.4 kb, 2.4 kb and 3.0 kb of 5’ flank sequence were used to generate lacZ fusion constructs for cpr-3, cpr-4, cpr-5 and cpr-6 respectively. However, it should be noted that 700 bp of the 3.0 kb fragment used for the cpr-6::lacZ construct are derived from the original pJB8 cosmid vector. Since this 700bp region is located at the most distal end of the 3.0 kb fragment (Chapter 3, Section 3.2.12), there is a maximum of 2.3 kb of 5’ flank sequence from the cpr-6 gene available for use with lacZ fusion constructs. Of this 2.3 kb, at least 1.35 kb of 5’ flank sequence most proximal to the cpr-6 gene are thought to be contiguous with the C.elegans genome from diagnostic digests of C.elegans genomic DNA (Chapter 3, Section 3.2.12).

The second consideration was the amount of coding sequence to include in each construct. The cathepsin B-like enzymes encoded by cpr-3, cpr-4, cpr-5 and cpr-6 all possess putative signal peptides. Therefore, it was possible that the use of too much coding sequence might result in the amino terminal prepeptide overriding the nuclear localisation signal of the lacZ fusion vector. Thus, I decided to use the minimum amount of coding sequence, resulting in no more than eight amino acid residues of the endogenous gene being coded for in the lacZ fusion. Three genes, cpr-3, cpr-4 and cpr-6, possess introns either immediately after or very close to the ATG translation initiation codon. For these three genes, the first intron was also included in the fusion transgene. These introns were included because potential regulatory elements have previously been observed in the first intron of cpr-1 (Ray and McKerrow, 1992). Though no such regulatory elements were identified in the first introns of cpr-3, cpr-4 and cpr-6, these introns were included as an added precaution.
5.3.6. Vector construction

A summary of the general approach used to obtain the fragments necessary to generate in-frame translational fusions of the 5' flank regions of each gene with the \textit{lacZ} gene of the pPD21.28 or pPD95.03 vectors is shown in Figure 5.18. For each gene, two PCR primers, X and Y were designed. Primer ‘X’ was designed to anneal upstream of the unique restriction site ‘U’ for each gene. These restriction sites (marked ‘U’) were identified in the sequenced region of the 5’ flank of each gene (Figure 5.19). The sites were subsequently shown to be unique within the cloned 5’ flanks of each gene by digestion of the subclones CL#034/1.7/KS+, ZK1055/2.8/KS+, W02B2/3/KS+ and C25B8/3.8/KS+, containing the 5’ flanks of \textit{cpr}-3, \textit{cpr}-4, \textit{cpr}-5 and \textit{cpr}-6 respectively (see Chapter 3, Figures 3.25, 3.26, 3.27 and 3.28), with the appropriate restriction endonuclease. Primer ‘Y’ was designed to link the coding region of each gene in frame with the \textit{lacZ} gene of the \textit{lacZ} fusion vectors. Therefore, for each gene, primer ‘Y’ was designed to anneal to the coding region downstream of the ATG translation initiation codon, and an appropriate restriction endonuclease site ‘A’ was engineered into the 5’ flank of the primer to facilitate cloning into the polylinker of the \textit{lacZ} fusion vectors. The PCR primers ‘X’ and ‘Y’ were used to amplify the intervening sequences for each gene, using the heat stable proof-reading Pwo DNA polymerase and a total of only fifteen cycles of PCR to minimise the number of mutations introduced into the amplified region. The sequences of the ‘X’ and ‘Y’ primers for each gene are given in (Chapter 2, Section 2.15). The sequences of the ‘Y’ primers are also shown in Figure 5.20, along with the region of each gene to which these primers anneal. The amplification products were then digested with the appropriate restriction endonucleases recognising sites ‘U’ and ‘A’ to generate 75-300bp fragments for linking the bulk of the 5’ flank of each gene to the \textit{lacZ} gene.

The bulk of the 5’ flank region of each gene was obtained by restriction endonuclease digestion of the subclones CL#034/1.7/KS+, ZK1055/2.8/KS+, W02B2/3/KS+ and C25B8/3.8/KS+, with the appropriate restriction endonucleases to release the cloned 5’ flank sequences. In all four cases, the subclones were digested with the restriction endonuclease that recognised the unique restriction site ‘U’ in the

251
sequenced region of the 5\' flank of each of the four genes (Figure 5.18). The subclones were digested at site 'B' (Figure 5.18), either with the restriction endonuclease originally used to subclone the inserts (CL#034/1.7/KS+, ZK1055/2.8/KS+ and C25B8/3.8/KS+) or with a restriction endonuclease present in the MCS of the pBluescript II KS- vector (W02B2/3/KS+). In the latter case, this was necessary to obtain a fragment that could be cloned into the polylinker of the lacZ fusion vector. The resulting fragments were gel purified prior to ligation into the lacZ fusion vectors. The restriction endonucleases which recognise sites 'A', 'B' and 'U' that were used in the preparation of these fragments are shown in Table 5.6A. The size of the fragments inserted into the lacZ fusion vectors are also shown in Table 5.6B.

5.3.6.1. Screening the PCR amplified linkers for mutations

The PCR amplified regions for cpr-3, cpr-4 and cpr-6 were digested with the restriction endonucleases that recognised sites 'A' and 'U' and subcloned into the pBluescript II SK- plasmid. The ligation products were transformed into *E. coli* XL1-blue and colonies were screened by colony PCR, using the M13 Forward and Reverse Primers which anneal to sequences flanking the polylinker of the plasmid vector. Plasmid DNA was subsequently prepared from colonies which generated PCR amplification products of the appropriate size. At least three clones for each gene were sequenced using the T3 and / or the T7 PCR primers. The nucleotide sequences of these clones were compared to the 5' flank sequences of the appropriate genes. This comparison revealed that at least one of the cloned PCR amplified regions for cpr-3, cpr-4 and cpr-6 shared complete sequence identity to the cloned 5' flanks of these genes, and therefore that these clones did not possess any PCR generated mutations. These clones were selected for further use. For cpr-5, the unique restriction site 'U' was recognised and cleaved by the restriction endonuclease BssHII. The pBluescript II SK- plasmid possesses two BssHII sites flanking the polylinker and therefore it was not possible to clone and sequence the PCR amplified linkers for cpr-5.
5.3.6.2. Preparation of the lacZ reporter constructs

The approach used to generate all the lacZ reporter constructs is summarised in Figure 5.21. In all cases, Fragment 1 of cpr-3, cpr-4, cpr-5 and cpr-6, previously prepared from subclones CL#034/1.7/KS+, ZK1055/2.8/KS+, W02B2/3/KS+ and C25B8/3.8/KS+ respectively (see Figure 5.18), were co-ligated into the polylinker of the lacZ fusion vectors with the appropriate PCR amplified linker region (Fragment 2). For cpr-3, cpr-4 and cpr-6, the PCR amplified linker regions (Fragment 2) were released from pBluescript II SK- using the restriction endonucleases that recognise sites ‘A’ and ‘U’ (Figure 5.18) and were gel purified. The PCR amplified linker region (Fragment 2) was co-ligated into the polylinker of the lacZ fusion vector with the appropriate 5’ flank DNA fragment (Fragment 1). For these three cases, the same restriction endonucleases that recognised sites ‘A’ and ‘B’ of the fragments to be cloned were also used to digest the polylinker of the lacZ fusion vector. For cpr-5, a different approach was required since the PCR amplified linkers could not be cloned into pBluescript II SK- as an intermediate step and therefore had not been sequenced to check for mutations. Thus the PCR amplified products (Fragment 2) for this cloned gene were digested with the restriction endonucleases that recognise sites ‘A’ and ‘U’ (Figure 5.18) and were cloned directly into the polylinker of the pPD21.28 lacZ fusion vector, with the cloned 5’ flank of this gene (Fragment 1). The polylinker of pPD21.28 did not possess the XhoI site at ‘B’ required for ligation to the cloned 5’ flank of cpr-5 (Fragment 1) and was therefore digested with SalI which produces ends compatible with those generated by XhoI.

5.3.6.3. Checking the integrity of the lacZ reporter constructs

E.coli XL1-blue, transformed with the products of the ligations described above, were plated onto L-agar supplemented with 100µg/ml Ampicillin and the resulting transformants screened using colony PCR. The colony PCR used two primers which annealed to the lacZ fusion vector sequences flanking the polylinker, the M13 Reverse primer and the lacZI2 primer or the SYNINT primer (the relative position and orientation of each of these primers is shown schematically in Figure 5.21). The nucleotide sequences of these primers are given in Chapter 2, Section 2.11.3.3. This
screen helped determine which clones contained the correct fragments and the correct number of fragments ligated into the lacZ fusion vectors. However, the sizes of the PCR products produced were sufficiently large in some cases, to make it hard to determine how many copies of the small PCR linker had also been subcloned into the vectors. Therefore, colonies that generated PCR products of the correct size were further analysed. Clones containing the lacZ fusion constructs for cpr-3, cpr-4 and cpr-6 were analysed by repeating the colony PCR but using the PCR primer ‘X’ for each gene, instead of the M13 Reverse primer. This would reveal whether the 5' flank DNA fragment was ligated into the vector in the correct orientation with a PCR linker of the correct size, since a PCR product would only be generated if the 5' flank fragment was in the correct orientation and the size of the product would be dependent on the size of the PCR linker. For cpr-6 the NUCLOC primer was also used instead of the lacZ/2 primer because it was closer to the polylinker. The position and orientation of the NUCLOC primer is shown schematically in Figure 5.21. The PCR linker for cpr-6 is only 75 bp and therefore I decided that using a primer more proximal to the polylinker than lacZ/2 would help to more accurately determine the size of the PCR linker. Together, the colony PCR results indicated that several clones for each of these constructs possessed a single copy of the 5' flank DNA fragment in the correct orientation and a single copy of the correct PCR linker cloned into the lacZ fusion vector. This approach was not used to screen the clones isolated for the lacZ fusion with cpr-5 because regions of these clones would require sequencing (since the PCR linkers had not been sequenced at an intermediate stage). Thus plasmid DNA was prepared from three colonies which generated PCR products of the correct size in the original colony PCR screen. The DNA was sequenced using primer ‘X’ for cpr-5, the SYNINT primer and the M13 Reverse primer. The sequencing, together with the initial colony PCR data, revealed that one of the selected clones contained a single copy of the correct 5' flank fragment and a single copy of the correct PCR linker, with no PCR generated mutations, all ligated in the correct orientation into the pPD21.28 lacZ fusion vector. Furthermore, the sequencing revealed that no mutations had been introduced during the ligation step. Since the selected clones for the cpr-3, cpr-4 and cpr-6 lacZ fusion constructs were not sequenced during this stage, plasmid DNA was prepared from overnight cultures of these and digested with the restriction endonucleases that recognised sites ‘A’, ‘B’ and ‘U’ to
ensure that the ligations had not resulted in mutations at these sites. The results of this indicated that all the restriction endonuclease sites utilised in the cloning strategies could be recut and therefore were not mutated.

5.3.6.4. Final analysis of the lacZ reporter constructs prior to injection

Medium or large scale plasmid DNA preparations were made from the chosen clones containing the constructs for each of the four genes, cpr-3, cpr-4, cpr-5 and cpr-6. A sample of each construct was subjected to a single round of sequencing using either the NUCLOC or SYNINT primers. This sequencing confirmed that the restriction enzyme site ‘A’ for each gene was intact and that translational gene fusions had been generated (Figure 5.22). The results of the previous colony PCR screens, diagnostic digests and sequencing reactions indicated that the generation of the reporter constructs had been successful, with the correct number and size of fragments being cloned into each of the lacZ fusion vectors, and no evidence of the introduction of mutations during these steps. The chosen constructs containing the 5’ flanks of cpr-3, cpr-4 and cpr-6 cloned into pPD95.03 were named cpr-3::lacZ, cpr-4::lacZ and cpr-6::lacZ respectively. The chosen construct containing the 5’ flank of cpr-5 cloned into pPD21.28 was named cpr-5::lacZ. The constructs cpr-3::lacZ, cpr-4::lacZ, cpr-5::lacZ and cpr-6::lacZ, containing 1.4kb, 2.4kb, 2.4kb and 3.0kb of 5’ flank sequence of the appropriate gene respectively, were subsequently used to generate transgenic lines of C.elegans.

5.3.7. Generating C.elegans lines transgenic for the reporter constructs and staining for β-Galactosidase activity

The syncytial Gonads of N2 wild type young adult hermaphrodite worms were microinjected with 200ng/μl of the desired construct and 100ng/μl of the pRF4 marker plasmid, containing the dominant rol-6(su1006) collagen mutation (Mello et al., 1991). This mutation results in a helically twisted cuticle which causes the worm to roll over about its longitudinal axis and to move in circles (the roller phenotype). F1 progeny displaying the roller phenotype were selected and 10 - 20 of these animals seeded onto each of several fresh plates. Single F2 progeny displaying the roller phenotype were
subsequently selected from these plates and seeded onto fresh plates. F2 progeny yielding 10% - 90% offspring with the roller phenotype were selected as transmitting lines. The microinjections produced four, five, six and two transmitting lines for the constructs *cpr-3::lacZ*, *cpr-4::lacZ*, *cpr-5::lacZ* and *cpr-6::lacZ*, respectively.

For staining, 10-20 larval and adult worms with the roller phenotype were seeded onto fresh plates and cultured until the bacterial lawn was almost cleared. The worms were then fixed with methanol and acetone and stained for β-galactosidase activity using X-gal as a substrate (Chapter 2, Sections 2.17.4 and 2.17.5).

5.3.8. Gut-specific expression of *cpr-3, cpr-4, cpr-5 and cpr-6*

The tissue-restricted patterns of expression generated by the four reporter constructs *cpr-3::lacZ*, *cpr-4::lacZ*, *cpr-5::lacZ* and *cpr-6::lacZ* are all remarkably similar (Figures 5.23, 5.24, 5.25 and 5.26). Furthermore, for all four constructs, this tissue-restricted pattern of expression was reproducible between different lines generated from the same construct. In all cases, large stained nuclei were localised along either side of the intestine, posteriorly of the pharyngeal-intestinal valve and anteriorly of the anus.

The entire intestine is derived exclusively from a single founder cell, the E cell, which gives rise to no other tissue. The daughter cells of E are the first to enter the body cavity during gastrulation, 90 minutes after the first cleavage. By 300 minutes, these cells have formed two rows, either side of the developing gut. The anterior most cells of each row subsequently divide dorso-ventrally to give rise to the four cells of int1, which attaches anteriorly to the pharyngeal-intestinal valve. The remaining intestinal cells undergo a 90° left hand twist (Sulston *et al.*, 1983). This twist subsequently increases to 180° and is thought to be a result of packing of the posterior nuclei. A schematic diagram of part of the intestinal tract, demonstrating these features is shown in Figure 5.27. Thus, by hatch, the intestine is comprised of a hollow tube, made up of twenty mononucleate cells which can be divided into nine units along its length (int1 to int9). The lumen of the intestine is surrounded by the four cells of int1 at the anterior end, and subsequently by pairs of cells that make up int2 to int9. Fifteen minutes after the beginning of the L1 lethargis, the nuclei of the intestinal cells divide. Normally the six
anterior most nuclei (of int1 and int2) do not divide, and any of the four posterior most nuclei (of int8 and int9) may also fail to divide. Thus between 10 and 14 of the intestinal nuclei divide, without cytokinesis, to produce between 30 and 34 gut nuclei in twenty cells (Sulston and Horvitz, 1977). All the nuclei of the intestine, including those that do not demonstrate nuclear division, undergo an endoreduplication event during lethargis of each larval moult. This event occurs without chromosome condensation and results in all the cells of the gut having a ploidy of 32C by the final moult (Hedgecock and White, 1985).

The features of the gut nuclei described above make the task of distinguishing the intestine from other tissues relatively simple. Thus the position, distribution and large size of the staining nuclei obtained with all the transgenic lines generated by the four reporter constructs, cpr-3::lacZ, cpr-4::lacZ, cpr-5::lacZ and cpr-6::lacZ indicate that they belong to intestinal cells (Figures 5.23, 5.24, 5.25 and 5.26, respectively). Furthermore, though individual intestinal cells could not be discerned, pairs of stained nuclei were frequently observed. These probably represent the binucleated cells of the intestine. These results suggest that expression of cpr-3, cpr-4, cpr-5 and cpr-6 is restricted specifically to the intestinal cells of C.elegans. Indeed, the expression patterns generated by cpr-3::lacZ, cpr-4::lacZ, cpr-5::lacZ and cpr-6::lacZ are very similar to those observed with lacZ fusions of two P-glycoprotein genes, pgp-1 and pgp-3, both of which show tissue-restricted expression in the intestinal cells of C.elegans (Lincke et al., 1993). At this point, it should be noted that the nuclear localisation of the β-galactosidase gene is a result of the nuclear localisation signal of the lacZ fusion vectors used, and therefore is not representative of the subcellular localisation of the products of the endogenous copies of cpr-3, cpr-4, cpr-5 and cpr-6.

5.3.8.1. Variable expression of the lacZ fusion transgenes in the intestine

For all the lines generated from the four lacZ constructs for cpr-3, cpr-4, cpr-5 and cpr-6, tissue-restricted staining of only the gut cell nuclei was consistent. However, there was a significant degree of variation in the number of stained nuclei, and the staining intensity. This variation was not only observed between lines generated from different constructs and between lines generated from the same constructs but also
between different animals from the same line. The type of variation observed for all four transgene fusions is consistent with mosaicism resulting from the transgenes being maintained as extrachromosomal arrays (discussed in Section 5.3.1). However, other factors could also account for the observed variation. First, the variation could be a result of fixation artefacts, with incomplete fixation resulting in less than the full complement of intestinal cells being equally available for staining. Second, the variation could be a result of the limits of sensitivity of the staining method used. Thus, some intestinal cells may not be expressing the fusion transgene at sufficient levels for detection.

5.3.8.2. Variability in the number and intensity of staining nuclei may be due to the different lacZ fusion vectors used

One of the most notable differences in the staining patterns observed with the different constructs was in sensitivity. Thus the gut cell nuclei of all the transgenic lines generated from constructs utilising the pPD95.03 lacZ fusion vector (cpr-3::lacZ, cpr-4::lacZ and cpr-6::lacZ) produced staining after 10min incubation with the X-gal substrate. In contrast, lines generated from constructs generated from the pPD21.28 lacZ fusion vector (cpr-5::lacZ) frequently took up to 72 hours to produce visibly stained nuclei using the same fixation and staining procedures, and the frequency of staining worms was also much lower. There was also a marked difference in the degree of mosaicism observed between lines generated using the pPD95.03 lacZ fusion constructs and those generated using the pPD21.28 vector. Up to 32 staining gut-cell nuclei were observed relatively frequently in larval and adult stages of lines containing the constructs cpr-3::lacZ, cpr-4::lacZ and cpr-6::lacZ. In contrast, no more than 20 staining gut-cell nuclei were observed for lines containing the construct cpr-5::lacZ. These observed differences may be a result of the different lacZ fusion vectors used to generate the different constructs. Thus all lines generated from different constructs utilising the same pPD95.03 vector produced more rapid and more frequent staining, in addition to lower levels of mosaicism, than lines derived from constructs utilising the pPD21.28 vector. The pPD95.03 vector is a member of the second generation of lacZ fusion vectors, produced by A.Fire, which contain introns in the lacZ coding region.
These vectors exhibit sensitivity two to three orders of magnitude greater than the older vectors such as pPD21.28 when assayed using the unc-54 promotor and also exhibit reduced levels of mosaicism with certain promotor sequences (A.Fire, pers.comm.). Therefore, the features of the new generation of lacZ fusion vectors could explain the differences observed. In turn, this suggests that the variation observed is not a result of 'true' mosaicism (i.e. loss of the transgene array in some intestinal cells) but rather a result of insufficient levels of expression of the transgene for detection in certain cells. Towards confirming this observation, it would be necessary to generate constructs with the same 5' flank sequences fused to different lacZ fusion vectors.

5.3.8.3. *cpr-3*, *cpr-4*, *cpr-5* and *cpr-6* may be expressed in different regions of the intestine

Despite the variation observed with the staining patterns for all four gene fusions, resulting from mosaicism or other factors, there was some evidence of differential expression of the four gene fusions within the intestine. For example, a comparison of the expression patterns obtained with the *cpr-3::lacZ* and *cpr-4::lacZ* transgene fusions reveals some interesting differences. For *cpr-3::lacZ*, the majority of intestinal cell nuclei exhibited strong staining for β-galactosidase activity at two distinct regions in adult worms, the anterior-most end of the intestine and the region spanning the mid-posterior end of the intestine (Figure 5.23A). For *cpr-4::lacZ*, the majority of adults exhibited equally strong staining of intestinal cell nuclei along the entire length of the gut (Figure 5.24A). For *cpr-3::lacZ*, there was also some evidence of a switch in the pattern of expression occurring during larval development. The majority of L₁ and L₂ larvae exhibited intense staining of intestinal cell nuclei spanning virtually the entire length of the intestine while the majority of L₃ and L₄ larvae exhibited the more polarised staining pattern observed with adult stages (Figure 5.23). In contrast, for *cpr-4::lacZ*, no staining pattern was more predominant than any other for the larval stages.

These observations suggest that it is possible that the cathepsin B-like genes of *C.elegans* may be differentially expressed along the length of the intestine and, in some cases, may be expressed in different cells of the gut during *C.elegans* development. It is worth noting that P.MacMorris and T.Blumenthal (pers.comm.) have observed evidence
of a conserved pattern in the appearance of gut cell nuclei expressing the vit-2 gene in adult worms, using lacZ gene fusions. Initially, nuclei of the mid and posterior gut were seen to express vit-2, followed by additional expression at a third focus at the anterior of the gut. Later, expression extended from each focus to include the surrounding cells until all nuclei of int2 to int8 stained. However, as a result of the mosaic-like variability observed for the gene fusions reported here, and the fact that no one expression pattern was exclusively associated with a given lacZ fusion construct, any conclusions regarding expression patterns restricted to certain regions of the gut, or patterns which vary with the development of the worm, can only be made with caution.

5.3.9. The temporal expression patterns indicated from the lacZ fusion analysis correlate with those indicated by s-q rtPCR

As discussed in Section 5.3.1, all the lacZ fusion constructs are maintained as extrachromosomal arrays in the transgenic worms. Since the arrays are not linked to chromosomes, they may not segregate properly during mitosis or meiosis. This results in transgenic worms not only exhibiting mosaicism in somatic tissues but also transmitting the transgenes in a non-Mendelian manner. Thus, it was not possible to accurately score the proportion of worms that stained for each developmental stage, since there was no way of determining whether lack of staining was due to lack of expression from an array or due to absence of an array. Accordingly, for each gene, accurate comparisons between the relative transcript abundance (determined by s-q rtPCR analysis) and the relative frequency with which worms expressed the appropriate transgene at different stages of development could not be made. Nevertheless, surveys of stained populations of worms did reveal several strong trends that did not require such accurate analysis and these trends are in agreement with the temporal expression patterns of the four genes determined using s-q rtPCR analysis.

In essence, the s-q rtPCR data suggest that induction of expression of cpr-3, cpr-4 and cpr-5 occurs around the time of embryo hatch. The results obtained with the lacZ fusions suggest that induction of expression of cpr-3, cpr-4 and cpr-5 may occur during late embryogenesis rather than at hatch. This apparent discrepancy is probably due to the use of a mixed stage embryo RNA sample with the s-q rtPCR (and Northern blot)
analyses. The transcript abundance of cpr-3, cpr-4 and cpr-5 would be underrepresented in such a sample if these genes are only expressed in the later stages of embryogenesis, as suggested by the lacZ data. Such an under-representation would give rise to an artificially perceived increase in relative transcript abundance between the embryo and L₁ stages for each of these genes using s-q rTPCR and would result in the conclusion that induction of expression of these three genes was occurring at the time of hatch rather than during late embryogenesis.

The s-q rTPCR results suggest that cpr-3, cpr-4 and cpr-5 are all expressed during the four larval and adult stages of C.elegans development, albeit with different patterns of relative transcript abundance (Section 5.2.3.5). C.elegans strains transgenic for the lacZ fusions of cpr-3, cpr-4 and cpr-5 all stained for β-galactosidase activity during the four larval and adult stages of C.elegans development and therefore support the results obtained from the s-q rTPCR experiments.

The s-q rTPCR data for cpr-6 produced a temporal pattern of expression very different from the other three genes, with significant levels of expression first occurring around the L₃ stage and increasing through the subsequent L₄ and early adult stages. Lines transgenic for cpr-6::lacZ showed almost exactly the same pattern of expression as predicted by the s-q rTPCR, the principle difference being that the lacZ fusion data suggested that cpr-6 expression was being induced during the L₂ stage, slightly earlier than predicted from the graphical representation of the s-q rTPCR data. Thus the intestinal cell nuclei of L₃, L₄ and adult stages stained at a high frequency while those of L₂ stage larvae stained only at a low frequency and neither the L₁ or embryo stages stained at all.

These results demonstrate that both the s-q rTPCR and lacZ gene fusion approaches yield similar temporal patterns of expression for each of the four genes. Most notably, both approaches are capable of distinguishing the very different temporal pattern of expression of cpr-6 from those of the other three genes.

5.4. Discussion

The experiments described in this chapter were performed to determine the temporal and spatial patterns of expression of cpr-3, cpr-4, cpr-5 and cpr-6. Obtaining
information on when and where a gene is expressed can be very helpful in assigning potential biological roles for that gene’s product. When such analysis is performed on a family of related genes, the information obtained may also help to determine how the members of the gene family interact at the biological level.

5.4.1. Overview of results obtained

The Northern blot studies indicate that transcription results in a single predominant mRNA species for each of the genes cpr-3, cpr-4, cpr-5 and cpr-6, suggesting that these genes are not alternatively spliced. These results agree with the 5' RACE experiments (Chapter 4, Section 4.2.6) which indicate that transcription initiation occurs predominantly from one region of the 5' flank of each of the four genes. The Northern blot studies also suggest that expression of cpr-4, cpr-5 and cpr-6 is induced at some point at or after embryo hatch (Section 5.2.2), supporting the data obtained for these three genes using s-q rtPCR (Section 5.2.4). Such conclusions cannot be made for cpr-3 because the probe used to control for equal loading of RNA during Northern analysis failed to generate any signal.

The s-q rtPCR experiments suggest that cpr-3 is essentially constitutively expressed during the larval and adult stages of C.elegans development, but that expression of this gene may be increased slightly during progression from the embryo to L1 stage. For cpr-4, the results suggest that expression of this gene occurs predominantly during the larval stages of development, since transcript abundance is reduced both in the embryo and adult stages. The s-q rtPCR results suggest that cpr-5 expression is induced at hatch, rises to a maximum at the L2 stage and shows a slight decline in abundance in subsequent larval and adult stages. For cpr-6, the results suggest that expression of this gene remains low during the embryo and early larval stages of development, but is induced during the L3 stage and continues to increase through the L4 and early adult stages. In addition, both cpr-5 and cpr-6 exhibit highly developmentally regulated patterns of expression, showing over 15 fold increases in their transcript abundance between minimal and maximal expression. Finally, the temporal expression patterns of the four genes obtained using s-q rtPCR correlate well with the expression data obtained from the appropriate lacZ transgene fusion. Most notably, both
approaches identified *cpr-6* as having a very different temporal expression pattern from *cpr-3, cpr-4* and *cpr-5* and both approaches generated a very similar expression pattern for this gene. Finally, the results obtained using the *lacZ* transgene fusions suggest that all four genes are expressed exclusively in the intestinal cells of *C.elegans*.

5.4.2. Efficacy of the results obtained using *lacZ* reporter gene fusions

The extensive use of *lacZ* gene fusions has demonstrated that this approach is frequently capable of faithfully reproducing the spatial expression patterns of endogenous genes in *C.elegans*. However, this approach, like any other, may produce spurious results in some cases (discussed in Section 5.3.3) and therefore cannot be used as the sole means of determining the physiological expression patterns of endogenous genes. Thus, it will be necessary to confirm the results reported here by another approach, such as *in situ* hybridisation or antibody staining of whole worms. Though, the spatial expression patterns of the *cpr-3::lacZ, cpr-4::lacZ, cpr-5::lacZ* and *cpr-6::lacZ* gene fusions have not as yet been confirmed using an alternative method, there is a substantial amount of indirect evidence to suggest that they do reflect the spatial expression patterns of the endogenous genes.

All four *lacZ* gene fusions produced similar gut-restricted patterns of expression (Section 5.3.8). These results suggest that it is the presence of regulatory elements in the 5′ flanks of these four genes that is responsible for generating the expression patterns observed, since it is unlikely that four different 5′ flank sequences would reproducibly generate the same artefact. Furthermore, these patterns of expression were obtained using two different *lacZ* fusion vectors, pPD95.03 and pPD21.28 which suggests that the gut-restricted expression patterns observed are not a result of the type of vector used.

Ectopic expression generally results in a few additional cells staining in tissues outwith those which express the endogenous gene. Therefore, in cases where ectopic expression is occurring, one expects to observe staining in more than one tissue. The ability of all four *lacZ* gene fusions reported here to produce staining only within the intestinal cells therefore argues against ectopic expression occurring.
The temporal expression patterns of the four genes obtained using both s-q rtPCR and lacZ fusion transgene approaches correlate well (Section 5.3.9). The ability of two very different approaches to generate similar temporal expression patterns for each of the four genes provides good evidence in itself for the efficacy of both these approaches. The ability of all four lacZ fusion transgenes to reflect the temporal expression patterns of their endogenous genes suggests that all the regulatory elements required to drive appropriate temporal expression of these genes are present in the lacZ fusion constructs. In turn, these results suggest that the elements required to drive appropriate spatial expression of the four genes are also present within each construct. Accordingly, one would not expect to be able to reproduce the correct temporal expression pattern of a gene if it were being expressed in inappropriate tissues.

It is commonly recognised that the 3' UTRs of some mature mRNA transcripts may be important for regulating the stability of those transcripts (Chapter 4, Section 4.3.4.1). This may affect the expression pattern of a gene by stabilising the transcript in certain tissues or at certain times during development. None of the four gene fusions were made using their 3' flank sequences, but instead the unc-54 3' end was used. Therefore it is possible that the temporal and/or spatial expression patterns of the lacZ fusion transgenes might be affected by the lack of appropriate 3' flank sequence. However, again the correlation between the temporal expression patterns obtained using the s-q rtPCR and lacZ fusion approaches argue against this. The s-q rtPCR approach provided information on the relevant transcript abundance of each of the four genes from all tissues. These results correlated with the frequency with which the intestinal cells from different stages of C.elegans development exhibit staining, in lines transgenic for the appropriate construct. Thus, one would expect any alteration in the temporal and/or spatial expression patterns caused by the use of a different 3' end in the lacZ gene fusions to reduce the correlation observed with the s-q rtPCR approach.

One of the strongest pieces of evidence for the efficacy of the lacZ fusion results comes from data obtained previously for cpr-1 (Ray and McKerrow, 1992). The authors demonstrated that expression of this cathepsin B-like gene occurs exclusively in the gut using in situ hybridisation. More recently, this expression pattern has been confirmed using a lacZ gene fusion containing only the 5' flank sequences of cpr-1 (C.Britton, pers. comm.). Such data suggests that cathepsin B-like genes are localised to the intestinal
cells of *C. elegans*. Interestingly, the data also suggest that the absence of a 3' flank sequence in the *cpr-1::lacZ* gene fusion does not result in alteration of the tissues where the gene is expressed, suggesting that the 3' UTR of *cpr-1* may not be necessary for appropriate spatial expression of this gene. There is also circumstantial evidence for expression of *cpr-3, cpr-4, cpr-5* and *cpr-6* in the gut of *C. elegans*. Analysis of the 5' flank sequences of *vit-2* (MacMorris *et al.*, 1992) and *ges-1* (Egan *et al.*, 1995) have identified GATA motifs, including the VPE2 sequences, which are essential for the appropriate gut-specific expression of these two genes in *C. elegans*. All four genes analysed here, in addition to *cpr-1*, possess a number of sequences in their 5' flanks with homology to these GATA motifs, including VPE2 sequences (Chapter 4, Section 4.3.4.2), suggesting that these sequences may be responsible for the gut-specific expression of *cpr-1, cpr-3, cpr-4, cpr-5* and *cpr-6*.

The final evidence for gut-specific expression of *cpr-3, cpr-4, cpr-5* and *cpr-6* comes from the potential roles one might expect these genes to perform. Such roles are consistent with gut-specific expression of these genes and are discussed in the next section. Before discussing these roles, it should be noted that expression of *cpr-3, cpr-4, cpr-5* and *cpr-6* is not synonymous with activity of their enzymes since the DNA sequence data suggest that the enzymes encoded by these genes are translated as precursors which must be processed. Furthermore, endogenous enzyme inhibitors may further restrict the spatial and temporal distribution of active enzyme translated from each of these genes. For instance, though the *lacZ* transgene fusion data suggest that *cpr-3, cpr-4, cpr-5* and *cpr-6* are all expressed in the intestinal cells, the intestinal cells may possess different inhibitors resulting in different intestinal cells possessing distinct active cathepsin B-like enzymes.

The analysis of the temporal and spatial expression patterns of *cpr-3, cpr-4, cpr-5* and *cpr-6* suggests that these genes are highly regulated at the transcriptional level. This suggests that the distribution of the enzymes encoded by these genes will be primarily regulated at the transcriptional level. Accordingly, endogenous inhibitors and enzyme processing events are unlikely to result in significant differences between the expression pattern of the gene and the distribution of its active cathepsin B-like enzyme, but may fine tune the distribution of active enzyme once expressed. Such fine tuning may well be at the cellular level, possibly with some active enzymes being targeted to the lysosomes.
and others being secreted. Indeed, the cathepsin B-like enzyme encoded by \textit{cpr-5} does not possess an N-glycosylation signal suggesting that it may not be targeted to the lysosomes but secreted instead (Chapter 4, Section 4.3.3.2).

5.4.3. Potential biological roles for \textit{cpr-3}, \textit{cpr-4}, \textit{cpr-5} and \textit{cpr-6}

The data obtained from the \textit{lacZ} fusions of \textit{cpr-3}, \textit{cpr-4} and \textit{cpr-5} suggest that these genes are expressed exclusively in the intestinal cells of \textit{C.elegans} (Section 5.3.8). These data, in conjunction with the s-q rtPCR data (Section 5.3.9), also suggest that expression of these genes is induced prior to hatch during the threefold stage of embryogenesis and that expression continues in all subsequent larval and adult stages of development, albeit at different levels. Such a temporal and spatial expression profile is consistent with these genes having a role in digestion in the gut of \textit{C.elegans}. Vertebrate cathepsin B is thought to play an important role in protein turnover and degradation within the lysosomes (Chapter 1, Section 1.7.1) but may also play an important role as a precursor processing enzyme (Chapter 1, Section 1.7.2). Since immunohistochemical studies have detected cathepsin B in many human tissues including the epithelia of the gastro-intestinal tract (Howie \textit{et al.}, 1985), it is possible that this enzyme performs an important, but indirect, function in digestion (possibly by processing proteins involved in digestion or by regulating the turnover of proteins involved in digestion). Vertebrate cathepsin B is not secreted into the vertebrate gut lumen and thus is unlikely to perform a direct role in digestion. Therefore, it is also possible that the cathepsin B-like enzymes of \textit{C.elegans} may play an analogous indirect role in digestion within the intestinal cells.

Alternatively, these enzymes may play a more direct role in digestion. Several parasitic nematode and trematode species have been shown to release cysteine protease enzymes into the media when cultured \textit{in vitro} (Chapter 1, Section 1.8) and these enzymes are thought to represent excreted or secreted products. Such studies have suggested that two parasitic trematode species, \textit{F.hepatica} and \textit{H.cylindracea}, excrete/secrete cathepsin B-like enzymes into the media when cultured \textit{in vitro}. Indeed, active cathepsin B-like activity has been identified in the secretory granules of intestinal cells of \textit{F.hepatica} (Yamasaki \textit{et al.}, 1992). More evidence of a direct role for cathepsin B-like enzymes in digestion comes from the trematode \textit{S.mansoni}. \textit{S.mansoni} uses
haemoglobin digestion as the principle mechanism for obtaining essential amino acids and the major haemoglobinase of this trematode species (Sm31) has a cathepsin B-like activity (Gotz and Klinkert, 1993). Together, these data suggest that cathepsin B-like enzymes may be secreted into the intestinal lumen and therefore may be directly involved in digestion in nematode and trematode species. To date, cathepsin B-like multigene families have been identified in the parasitic nematode species, *H. contortus* and *O. ostertagi* (Chapter 1, section 1.9) in addition to the multigene family of *C. elegans* reported in this thesis. There is also some evidence for the presence of such a multigene family in the parasitic nematode species, *A. caninum* and the parasitic trematode species, *F. hepatica* (Chapter 1, Sections 1.9). These data suggest that cathepsin B-like multigene families may be common to many nematode and trematode species. This is in agreement with a direct role in digestion, since amplification of this gene family may have been required to produce several enzymes which, when acting together, can degrade the broad spectrum of substrates necessary for a direct role in digestion.

The *cpr-6* gene is also expressed in the intestinal cells of *C. elegans* and therefore may also perform a digestive role in conjunction with *cpr-3*, *cpr-4* and *cpr-5*. However, the temporal expression pattern of this gene suggests that induction of expression does not occur until the later larval stages of development. Such an expression pattern is not expected for a gene involved in digestion since early larval stages must also feed. This does not preclude *cpr-6* having a role in digestion but merely indicates that this gene is not required for digestion in early larval stages. Though *cpr-6* may perform a role in digestion, the unusual temporal expression pattern of this gene compared to *cpr-3*, *cpr-4* and *cpr-5* suggests that it may perform other related roles.

The primary function of the intestine is probably to secrete digestive enzymes and absorb the processed nutrients. However, the intestine of *C. elegans* also represents one of the main storage organs of the body, possessing numerous different storage granules in the cytoplasm of intestinal cells. These storage granules are thought to contain lipids, proteins and carbohydrates (White, 1988). Thus, the cathepsin B-like enzyme encoded by the *cpr-6* gene may be involved in turnover and degradation of stored proteins. Alternatively, this enzyme may be involved in processing stored precursor proteins required during later stages of development. The yolk proteins represent one such family of proteins which may require processing during later stages of *C. elegans* development.
These proteins are synthesised in the intestine of *C.elegans* (Kimble and Sharrock, 1983). The genes encoding these proteins are expressed exclusively in the intestinal cells of hermaphrodite worms, with expression starting at the late L4 stage and continuing throughout the lifespan of the adult hermaphrodite (Blumenthal et al., 1984). Since the yolk proteins of other organisms are synthesised as precursor proteins that require subsequent processing (Tata, 1976), the yolk proteins of *C.elegans* may also require processing in the intestine of adult hermaphrodite worms. *cpr-6* shows the appropriate temporal and spatial expression pattern for a potential role in yolk protein processing, being expressed in the intestine in the later larval stages of development prior to yolk protein gene expression. It is therefore tempting to speculate that this gene may be required for yolk protein processing in hermaphrodite worms. Since yolk proteins are expressed in a sex specific manner, the enzymes required to process these proteins may also be expressed in a sex specific manner. Therefore, as a first step towards investigating the possibility of *cpr-6* being involved in yolk protein processing, it would be useful to determine whether or not *cpr-6* is expressed in male worms. Since, if *cpr-6* was also found to be expressed exclusively in the intestinal cells of hermaphrodite worms, such findings would provide good circumstantial evidence of a role for this enzyme in yolk protein processing.

5.4.4. Biological interaction

The data reported in this chapter suggest that all four genes are expressed in the same tissue and possess overlapping patterns of temporal expression. These data suggest that the members of this gene family may encode proteins which act together to effect their biological roles. However, though the temporal expression patterns overlap, they are clearly different. This may result in different combinations of active cathepsin B-like enzymes at different stages of development. The possibility that *cpr-3, cpr-4, cpr-5* and *cpr-6* encode cathepsin B-like enzymes with different activities and substrate specificities (Chapter 4, Section 4.3.3.1) may be important in this respect since different combinations of enzymes may degrade or process very different combinations of protein substrates. Thus biological interaction between the products of this gene family could be very complex, with different enzymes interacting at different stages, possibly to effect
different but related functions. For example, one enzyme may be capable of performing both digestive and processing roles but the type of role performed may be dependent on interaction with different members of the gene family. Accordingly, the possible digestive roles for cpr-3, cpr-4 and cpr-5 and the possible processing role for cpr-6 are not mutually exclusive, since some or all of these enzymes may interact at various stages of development to effect different but related biological functions.

Since the data suggest that cpr-3, cpr-4, cpr-5 and cpr-6 may be expressed simultaneously in the same tissue at certain stages of C.elegans development, it is possible that some members of this gene family may be functionally redundant. It is unlikely that any one of the four genes reported here, or cpr-1, will be completely functionally redundant (i.e. encode an enzyme which shares identical activity and substrate specificity with another enzyme), since these five genes encode enzymes with diverged predicted amino acid sequences which may have different activities and substrate specificities. It is possible that additional, as yet unidentified, cathepsin B-like genes do exist in the C.elegans genome which have identical activities and substrate specificities to the enzymes encoded by the five genes isolated to date. However, this is unlikely since the Southern blot data presented in Chapter 3, Section 3.2.2 does not provide any evidence for the presence of additional cathepsin B-like genes with significant homology to cpr-3, cpr-4, cpr-5 and cpr-6 within the genome of C.elegans. Thus, if there is any functional redundancy, it is most likely to be a result of the ability of a combination of cathepsin B-like enzymes to mimic the activity and substrate specificity of another individual enzyme. Accordingly, though individual members of the cathepsin B-like gene family may play important roles in C.elegans biology, null mutations in individual genes may not result in lethal phenotypes because other members of the gene family, acting together, may be able to assume the biological function of the mutated gene. The presence of functionally redundant members within the cathepsin B-like multigene family would suggest that this gene family is essential for C.elegans biology. In order to test such a hypothesis, it would be necessary to obtain C.elegans strains each possessing a null mutation of a different member of the cathepsin B-like gene family. These strains could then be crossed to produce multiply mutant strains. Analysis of the phenotypes of strains with mutations in different combinations of cathepsin B-like genes might then reveal which genes play biologically interchangeable roles.
Table 5.1

The table shows the radioactivity data obtained from s-q rtPCR analysis of cpr-3 using the modified protocol. See Figure 5.10 for a summary of the protocol.

The regions of the six filters corresponding to the PCR amplification products derived from the cpr-3 and ama-1 transcripts were excised using the autoradiographs shown in Figure 5.11 as templates. The amount of radioactivity (in counts per minute, cpm) from each band was measured twice by counting in scintillant. The average of the two counts was taken. The relative transcript abundance of cpr-3 at each developmental stage was calculated as a ratio of the average cpm obtained from the PCR products derived from the cpr-3 transcripts to the average cpm obtained from the PCR products derived from the ama-1 transcripts. The mean and standard deviation of the cpr-3 relative transcript abundance ratio at each developmental stage was calculated from the results obtained from three independent PCR reactions. Accordingly, the mean and standard deviation for each developmental stage was calculated from filters 1, 3 and 5 or from filters 2, 4 and 6.
<table>
<thead>
<tr>
<th>Filter 1</th>
<th>Filter 2</th>
<th>Filter 3</th>
<th>Filter 4</th>
<th>Filter 5</th>
<th>Filter 6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Embryo</td>
<td>Embryo</td>
<td>Embryo</td>
<td>Embryo</td>
<td>Embryo</td>
<td>Embryo</td>
</tr>
<tr>
<td>Cmp</td>
<td>ACP</td>
<td>ACP</td>
<td>ACP</td>
<td>ACP</td>
<td>ACP</td>
</tr>
</tbody>
</table>

**Mean and Standard Deviation**

<table>
<thead>
<tr>
<th>Filter 1</th>
<th>Filter 2</th>
<th>Filter 3</th>
<th>Filter 4</th>
<th>Filter 5</th>
<th>Filter 6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Embryo</td>
<td>Embryo</td>
<td>Embryo</td>
<td>Embryo</td>
<td>Embryo</td>
<td>Embryo</td>
</tr>
<tr>
<td>Cmp</td>
<td>ACP</td>
<td>ACP</td>
<td>ACP</td>
<td>ACP</td>
<td>ACP</td>
</tr>
</tbody>
</table>

**Mean and Standard Deviation**

<table>
<thead>
<tr>
<th>Filter 1</th>
<th>Filter 2</th>
<th>Filter 3</th>
<th>Filter 4</th>
<th>Filter 5</th>
<th>Filter 6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Embryo</td>
<td>Embryo</td>
<td>Embryo</td>
<td>Embryo</td>
<td>Embryo</td>
<td>Embryo</td>
</tr>
<tr>
<td>Cmp</td>
<td>ACP</td>
<td>ACP</td>
<td>ACP</td>
<td>ACP</td>
<td>ACP</td>
</tr>
</tbody>
</table>

**Mean and Standard Deviation**

<table>
<thead>
<tr>
<th>Filter 1</th>
<th>Filter 2</th>
<th>Filter 3</th>
<th>Filter 4</th>
<th>Filter 5</th>
<th>Filter 6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Embryo</td>
<td>Embryo</td>
<td>Embryo</td>
<td>Embryo</td>
<td>Embryo</td>
<td>Embryo</td>
</tr>
<tr>
<td>Cmp</td>
<td>ACP</td>
<td>ACP</td>
<td>ACP</td>
<td>ACP</td>
<td>ACP</td>
</tr>
</tbody>
</table>

**Mean and Standard Deviation**

<table>
<thead>
<tr>
<th>Filter 1</th>
<th>Filter 2</th>
<th>Filter 3</th>
<th>Filter 4</th>
<th>Filter 5</th>
<th>Filter 6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Embryo</td>
<td>Embryo</td>
<td>Embryo</td>
<td>Embryo</td>
<td>Embryo</td>
<td>Embryo</td>
</tr>
<tr>
<td>Cmp</td>
<td>ACP</td>
<td>ACP</td>
<td>ACP</td>
<td>ACP</td>
<td>ACP</td>
</tr>
</tbody>
</table>

**Mean and Standard Deviation**

<table>
<thead>
<tr>
<th>Filter 1</th>
<th>Filter 2</th>
<th>Filter 3</th>
<th>Filter 4</th>
<th>Filter 5</th>
<th>Filter 6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Embryo</td>
<td>Embryo</td>
<td>Embryo</td>
<td>Embryo</td>
<td>Embryo</td>
<td>Embryo</td>
</tr>
<tr>
<td>Cmp</td>
<td>ACP</td>
<td>ACP</td>
<td>ACP</td>
<td>ACP</td>
<td>ACP</td>
</tr>
</tbody>
</table>

**Mean and Standard Deviation**

<table>
<thead>
<tr>
<th>Filter 1</th>
<th>Filter 2</th>
<th>Filter 3</th>
<th>Filter 4</th>
<th>Filter 5</th>
<th>Filter 6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Embryo</td>
<td>Embryo</td>
<td>Embryo</td>
<td>Embryo</td>
<td>Embryo</td>
<td>Embryo</td>
</tr>
<tr>
<td>Cmp</td>
<td>ACP</td>
<td>ACP</td>
<td>ACP</td>
<td>ACP</td>
<td>ACP</td>
</tr>
</tbody>
</table>

**Mean and Standard Deviation**

<table>
<thead>
<tr>
<th>Filter 1</th>
<th>Filter 2</th>
<th>Filter 3</th>
<th>Filter 4</th>
<th>Filter 5</th>
<th>Filter 6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Embryo</td>
<td>Embryo</td>
<td>Embryo</td>
<td>Embryo</td>
<td>Embryo</td>
<td>Embryo</td>
</tr>
<tr>
<td>Cmp</td>
<td>ACP</td>
<td>ACP</td>
<td>ACP</td>
<td>ACP</td>
<td>ACP</td>
</tr>
</tbody>
</table>

**Mean and Standard Deviation**

<table>
<thead>
<tr>
<th>Filter 1</th>
<th>Filter 2</th>
<th>Filter 3</th>
<th>Filter 4</th>
<th>Filter 5</th>
<th>Filter 6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Embryo</td>
<td>Embryo</td>
<td>Embryo</td>
<td>Embryo</td>
<td>Embryo</td>
<td>Embryo</td>
</tr>
<tr>
<td>Cmp</td>
<td>ACP</td>
<td>ACP</td>
<td>ACP</td>
<td>ACP</td>
<td>ACP</td>
</tr>
</tbody>
</table>

**Mean and Standard Deviation**

<table>
<thead>
<tr>
<th>Filter 1</th>
<th>Filter 2</th>
<th>Filter 3</th>
<th>Filter 4</th>
<th>Filter 5</th>
<th>Filter 6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Embryo</td>
<td>Embryo</td>
<td>Embryo</td>
<td>Embryo</td>
<td>Embryo</td>
<td>Embryo</td>
</tr>
<tr>
<td>Cmp</td>
<td>ACP</td>
<td>ACP</td>
<td>ACP</td>
<td>ACP</td>
<td>ACP</td>
</tr>
</tbody>
</table>

**Mean and Standard Deviation**

<table>
<thead>
<tr>
<th>Filter 1</th>
<th>Filter 2</th>
<th>Filter 3</th>
<th>Filter 4</th>
<th>Filter 5</th>
<th>Filter 6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Embryo</td>
<td>Embryo</td>
<td>Embryo</td>
<td>Embryo</td>
<td>Embryo</td>
<td>Embryo</td>
</tr>
<tr>
<td>Cmp</td>
<td>ACP</td>
<td>ACP</td>
<td>ACP</td>
<td>ACP</td>
<td>ACP</td>
</tr>
</tbody>
</table>

**Mean and Standard Deviation**

<table>
<thead>
<tr>
<th>Filter 1</th>
<th>Filter 2</th>
<th>Filter 3</th>
<th>Filter 4</th>
<th>Filter 5</th>
<th>Filter 6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Embryo</td>
<td>Embryo</td>
<td>Embryo</td>
<td>Embryo</td>
<td>Embryo</td>
<td>Embryo</td>
</tr>
<tr>
<td>Cmp</td>
<td>ACP</td>
<td>ACP</td>
<td>ACP</td>
<td>ACP</td>
<td>ACP</td>
</tr>
</tbody>
</table>

**Mean and Standard Deviation**

<table>
<thead>
<tr>
<th>Filter 1</th>
<th>Filter 2</th>
<th>Filter 3</th>
<th>Filter 4</th>
<th>Filter 5</th>
<th>Filter 6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Embryo</td>
<td>Embryo</td>
<td>Embryo</td>
<td>Embryo</td>
<td>Embryo</td>
<td>Embryo</td>
</tr>
<tr>
<td>Cmp</td>
<td>ACP</td>
<td>ACP</td>
<td>ACP</td>
<td>ACP</td>
<td>ACP</td>
</tr>
</tbody>
</table>

**Mean and Standard Deviation**

<table>
<thead>
<tr>
<th>Filter 1</th>
<th>Filter 2</th>
<th>Filter 3</th>
<th>Filter 4</th>
<th>Filter 5</th>
<th>Filter 6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Embryo</td>
<td>Embryo</td>
<td>Embryo</td>
<td>Embryo</td>
<td>Embryo</td>
<td>Embryo</td>
</tr>
<tr>
<td>Cmp</td>
<td>ACP</td>
<td>ACP</td>
<td>ACP</td>
<td>ACP</td>
<td>ACP</td>
</tr>
</tbody>
</table>

**Mean and Standard Deviation**

<table>
<thead>
<tr>
<th>Filter 1</th>
<th>Filter 2</th>
<th>Filter 3</th>
<th>Filter 4</th>
<th>Filter 5</th>
<th>Filter 6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Embryo</td>
<td>Embryo</td>
<td>Embryo</td>
<td>Embryo</td>
<td>Embryo</td>
<td>Embryo</td>
</tr>
<tr>
<td>Cmp</td>
<td>ACP</td>
<td>ACP</td>
<td>ACP</td>
<td>ACP</td>
<td>ACP</td>
</tr>
</tbody>
</table>

**Mean and Standard Deviation**

<table>
<thead>
<tr>
<th>Filter 1</th>
<th>Filter 2</th>
<th>Filter 3</th>
<th>Filter 4</th>
<th>Filter 5</th>
<th>Filter 6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Embryo</td>
<td>Embryo</td>
<td>Embryo</td>
<td>Embryo</td>
<td>Embryo</td>
<td>Embryo</td>
</tr>
<tr>
<td>Cmp</td>
<td>ACP</td>
<td>ACP</td>
<td>ACP</td>
<td>ACP</td>
<td>ACP</td>
</tr>
</tbody>
</table>

**Mean and Standard Deviation**

<table>
<thead>
<tr>
<th>Filter 1</th>
<th>Filter 2</th>
<th>Filter 3</th>
<th>Filter 4</th>
<th>Filter 5</th>
<th>Filter 6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Embryo</td>
<td>Embryo</td>
<td>Embryo</td>
<td>Embryo</td>
<td>Embryo</td>
<td>Embryo</td>
</tr>
<tr>
<td>Cmp</td>
<td>ACP</td>
<td>ACP</td>
<td>ACP</td>
<td>ACP</td>
<td>ACP</td>
</tr>
</tbody>
</table>
Table 5.2

The table shows the radioactivity data obtained from s-q rtPCR analysis of cpr-4 using the modified protocol. See Figure 5.10 for a summary of the protocol.

The regions of the six filters corresponding to the PCR amplification products derived from the cpr-4 and ama-1 transcripts were excised using the autoradiographs shown in Figure 5.12 as templates. The amount of radioactivity (in counts per minute, cpm) from each band was measured twice by counting in scintillant. The average of the two counts was taken. The relative transcript abundance of cpr-4 at each developmental stage was calculated as a ratio of the average cpm obtained from the PCR products derived from the cpr-4 transcripts to the average cpm obtained from the PCR products derived from the ama-1 transcripts. The mean and standard deviation of the cpr-4 relative transcript abundance ratio at each developmental stage was calculated from the results obtained from three independent PCR reactions. Accordingly, the mean and standard deviation for each developmental stage was calculated from filters 1, 3 and 5 or from filters 2, 4 and 6.
Table 5.3

The table shows the radioactivity data obtained from s-q rtPCR analysis of *cpr-5* using the modified protocol. See Figure 5.10 for a summary of the protocol.

The regions of the six filters corresponding to the PCR amplification products derived from the *cpr-5* and *ama-1* transcripts were excised using the autoradiographs shown in Figure 5.13 as templates. The amount of radioactivity (in counts per minute, cpm) from each band was measured twice by counting in scintillant. The average of the two counts was taken. The relative transcript abundance of *cpr-5* at each developmental stage was calculated as a ratio of the average cpm obtained from the PCR products derived from the *cpr-5* transcripts to the average cpm obtained from the PCR products derived from the *ama-1* transcripts. The mean and standard deviation of the *cpr-5* relative transcript abundance ratio at each developmental stage was calculated from the results obtained from three independent PCR reactions. Accordingly, the mean and standard deviation for each developmental stage was calculated from filters 1, 3 and 5 or from filters 2, 4 and 6.
Table 5.4

The table shows the radioactivity data obtained from s-q rtPCR analysis of cpr-6 using the modified protocol. See Figure 5.10 for a summary of the protocol.

The regions of the six filters corresponding to the PCR amplification products derived from the cpr-6 and ama-1 transcripts were excised using the autoradiographs shown in Figure 5.14 as templates. However, the samples taken from filters 3 and 5 were lost and therefore only the data for filters 1, 2, 4 and 6 are shown. The amount of radioactivity (in counts per minute, cpm) from each band was measured twice by counting in scintillant. The average of the two counts was taken. The relative transcript abundance of cpr-6 at each developmental stage was calculated as a ratio of the average cpm obtained from the PCR products derived from the cpr-6 transcripts to the average cpm obtained from the PCR products derived from the ama-1 transcripts. The mean and standard deviation of the cpr-6 relative transcript abundance ratio at each developmental stage was calculated from the results obtained from three independent PCR reactions. Accordingly, the mean and standard deviation for each developmental stage was calculated from filters 2, 4 and 6.
<table>
<thead>
<tr>
<th>Filter 2</th>
<th>Filter 4</th>
<th>Filter 6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adult 1</td>
<td>Adult 1</td>
<td>Adult 1</td>
</tr>
<tr>
<td>L1</td>
<td>L1</td>
<td>L1</td>
</tr>
<tr>
<td>L2</td>
<td>L2</td>
<td>L2</td>
</tr>
<tr>
<td>L3</td>
<td>L3</td>
<td>L3</td>
</tr>
<tr>
<td>L4</td>
<td>L4</td>
<td>L4</td>
</tr>
<tr>
<td>Embryo</td>
<td>Embryo</td>
<td>Embryo</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Measure</th>
<th>Mean</th>
<th>St. Dev.</th>
</tr>
</thead>
<tbody>
<tr>
<td>cpm Am 1 DNA</td>
<td>12.43</td>
<td>6.34</td>
</tr>
<tr>
<td>cpm Am 1 PCR product</td>
<td>12.43</td>
<td>6.34</td>
</tr>
<tr>
<td>cpm Am 2 DNA</td>
<td>12.43</td>
<td>6.34</td>
</tr>
<tr>
<td>cpm Am 2 PCR product</td>
<td>12.43</td>
<td>6.34</td>
</tr>
<tr>
<td>Ratio Am 1:Am 2</td>
<td>1.00</td>
<td>1.00</td>
</tr>
</tbody>
</table>

Filter 1

<table>
<thead>
<tr>
<th>Measure</th>
<th>Mean</th>
<th>St. Dev.</th>
</tr>
</thead>
<tbody>
<tr>
<td>cpm Am 1 DNA</td>
<td>12.43</td>
<td>6.34</td>
</tr>
<tr>
<td>cpm Am 1 PCR product</td>
<td>12.43</td>
<td>6.34</td>
</tr>
<tr>
<td>cpm Am 2 DNA</td>
<td>12.43</td>
<td>6.34</td>
</tr>
<tr>
<td>cpm Am 2 PCR product</td>
<td>12.43</td>
<td>6.34</td>
</tr>
<tr>
<td>Ratio Am 1:Am 2</td>
<td>1.00</td>
<td>1.00</td>
</tr>
</tbody>
</table>

Filter 3

<table>
<thead>
<tr>
<th>Measure</th>
<th>Mean</th>
<th>St. Dev.</th>
</tr>
</thead>
<tbody>
<tr>
<td>cpm Am 1 DNA</td>
<td>12.43</td>
<td>6.34</td>
</tr>
<tr>
<td>cpm Am 1 PCR product</td>
<td>12.43</td>
<td>6.34</td>
</tr>
<tr>
<td>cpm Am 2 DNA</td>
<td>12.43</td>
<td>6.34</td>
</tr>
<tr>
<td>cpm Am 2 PCR product</td>
<td>12.43</td>
<td>6.34</td>
</tr>
<tr>
<td>Ratio Am 1:Am 2</td>
<td>1.00</td>
<td>1.00</td>
</tr>
</tbody>
</table>
Table 5.5

Statistical analysis of the temporal expression patterns of *cpr-3*, *cpr-4*, *cpr-5* and *cpr-6* (shown in Figure 5.16) using Student’s *t*-test. For each gene, the relative transcript abundance for each stage was compared to the next stage in development (i.e. embryo and L₁, L₁ and L₂, L₂ and L₃, L₃ and L₄, L₄ and adult). Each *t*-test was performed with three sets of paired data. The three sets of data were derived from the three independent PCR reactions (A, B and C, Figure 10) and each pair comprises the relative transcript abundance ratios of two adjacent developmental stages derived from the same PCR reaction. The probabilities were interpolated from a table of tabulated values of *t* using two degrees of freedom.

The *t*-test null hypothesis assumes that two samples have been derived from the same population, and therefore that any differences in the mean of these samples can be accounted for by sampling errors. Accordingly, the percentage probabilities indicate the probability of the relative transcript abundance ratios of two adjacent developmental stages being derived from the same population. Probabilities of less than or equal to 5% were deemed significant (S) and probabilities less than or equal to 1% were deemed highly significant (HS).
<table>
<thead>
<tr>
<th></th>
<th>cpr-3</th>
<th>cpr-4</th>
<th>cpr-5</th>
<th>cpr-6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Embryo-L₁</td>
<td>2.70% S</td>
<td>0.84% HS</td>
<td>1.13% S</td>
<td>7.81%</td>
</tr>
<tr>
<td>L₁-L₂</td>
<td>81.02%</td>
<td>9.39%</td>
<td>2.08% S</td>
<td>37.52%</td>
</tr>
<tr>
<td>L₂-L₃</td>
<td>41.72%</td>
<td>60.85%</td>
<td>4.16% S</td>
<td>79.78%</td>
</tr>
<tr>
<td>L₃-L₄</td>
<td>9.63%</td>
<td>10.94%</td>
<td>2.38% S</td>
<td>0.33% HS</td>
</tr>
<tr>
<td>L₄-Adult</td>
<td>11.54%</td>
<td>2.64% S</td>
<td>1.12% S</td>
<td>11.21%</td>
</tr>
</tbody>
</table>
Table 5.6A

The table shows the restriction endonucleases that recognise sites ‘A’, ‘B’ and ‘U’ (Figure 5.18) which were used in the generation of the lacZ fusion transgene constructs of cpr-3, cpr-4, cpr-5 and cpr-6.

Table 5.6B

The table shows the sizes of Fragment 1 and Fragment 2 (see Figure 5.18) after digestion with the restriction endonucleases which recognise sites ‘A’, ‘B’ and ‘U’. The total size of the 5’ flank used for each gene is also indicated.

Note: 700bp of the 3kb 5’ flank fragment used for cpr-6 is from the original pJB8 cosmid vector. Therefore, a maximum of 2.3kb of cpr-6 5’ flank sequence was used in the generation of the lacZ fusion transgene construct for this gene.
Table 5.6A

<table>
<thead>
<tr>
<th></th>
<th>enzyme used at site ‘A’</th>
<th>enzyme used at site ‘B’</th>
<th>enzyme used at site ‘U’</th>
</tr>
</thead>
<tbody>
<tr>
<td>cpr-3 (subclone CL#034/1.7/KS+)</td>
<td>BamHI</td>
<td>HindIII</td>
<td>EcoRV</td>
</tr>
<tr>
<td>cpr-4 (subclone ZK1055/2.8/KS+)</td>
<td>Xmal</td>
<td>HindIII</td>
<td>KpnI</td>
</tr>
<tr>
<td>cpr-5 (subclone W02B2/3/KS+)</td>
<td>Xmal</td>
<td>Xhol</td>
<td>BssHII</td>
</tr>
<tr>
<td>cpr-6 (subclone C25B8/3.8/KS+)</td>
<td>BamHI</td>
<td>PstI</td>
<td>SalI</td>
</tr>
</tbody>
</table>

Table 5.6B

<table>
<thead>
<tr>
<th></th>
<th>Size of Fragment 1 after digeston with enzymes recognising sites ‘B’ and ‘U’</th>
<th>Size of Fragment 2 after digestion with enzymes recognising sites ‘A’ and ‘U’</th>
<th>total size of 5’ flank used in the lacZ fusion constructs</th>
</tr>
</thead>
<tbody>
<tr>
<td>cpr-3 (subclone CL#034/1.7/KS+)</td>
<td>1.1 kb</td>
<td>300 bp</td>
<td>1.4 kb</td>
</tr>
<tr>
<td>cpr-4 (subclone ZK1055/2.8/KS+)</td>
<td>2.2 kb</td>
<td>210 bp</td>
<td>2.4 kb</td>
</tr>
<tr>
<td>cpr-5 (subclone W02B2/3/KS+)</td>
<td>2.2 kb</td>
<td>150 bp</td>
<td>2.4 kb</td>
</tr>
<tr>
<td>cpr-6 (subclone C25B8/3.8/KS+)</td>
<td>3.0 kb</td>
<td>75 bp</td>
<td>3.0 kb</td>
</tr>
</tbody>
</table>
The autoradiographic results of Northern blot analysis obtained after medium stringency washing. Northern blots were prepared from 4μg of total RNA isolated from mixed stage cultures of *C. elegans* (Lane 1) and from *C. elegans* embryos (Lane 2). The denaturing agarose gel used to generate the Northern blots is shown in Figure 5.3A.

A: Northern blot probed with the cDNA clone of *cpr-3* (cm12b6)
B: Northern blot probed with the cDNA clone of *cpr-4* (cm14e3)
C: Northern blot probed with the cDNA clone of *cpr-5* (cm04d10)
D: Northern blot probed with the cDNA clone of *cpr-6* (cm01a5)
Figure 5.2

The autoradiographic results of Northern blot analysis obtained after high stringency washing. Northern blots were prepared from 4μg of total RNA isolated from mixed stage cultures of *C.elegans* (Lane 1) and from *C.elegans* embryos (Lane 2). The denaturing agarose gel used to generate the Northern blots is shown in Figure 5.3A.

A: Northern blot probed with the cDNA clone of *cpr-4* (cm14e3)
B: Northern blot probed with the cDNA clone of *cpr-5* (cm04d10)
C: Northern blot probed with the cDNA clone of *cpr-6* (cm01a5)
Figure 5.3

A: An ethidium bromide stained 1.0% denaturing agarose gel showing four aliquots of 4μg of total RNA isolated from mixed stage cultures of *C. elegans* (Lane 1) and *C. elegans* embryos (Lane 2) prior to Northern blotting.

B: The autoradiographic results obtained after high stringency washing of two Northern blots hybridised with a 7kb *BamHI* rDNA fragment containing a single repeat of the 26s and 18s ribosomal DNA sequences. The Northern blots were generated from the denaturing agarose gel shown in A.
Figure 5.4

The autoradiographic results obtained from Northern blot analysis of poly(A)$^+$ selected RNA isolated from mixed stage cultures of *C. elegans* (Lane 1) and *C. elegans* embryos (Lane 2)

The Northern blot was hybridised with a riboprobe generated from the *cpr-3* cDNA clone (cm12b6) and washed at high stringency.
Figure 5.5

Autoradiographic results obtained from s-q rtPCR analysis of *cpr-3*

Each of 20 cDNA samples, generated from total RNA isolated at two hour intervals post hatch, was amplified by PCR. Each PCR reaction was performed using primers specific for *cpr-3* and the internal control gene, *ama-1*. The PCR reactions were electrophoresed through a 1.2% agarose gel, Southern blotted to Hybond N (Amersham International plc) and hybridised with the same primers used in the PCR reaction, 5’ end-labelled with ($\gamma^{32}$P)ATP. The Southern blots were washed at medium stringency.
Figure 5.6

 Autoradiographic results obtained from s-q rtPCR analysis of *cpr-4*

 Each of 20 cDNA samples, generated from total RNA isolated at two hour intervals post hatch, was amplified by PCR. Each PCR reaction was performed using primers specific for *cpr-4* and the internal control gene, *ama-1*. The PCR reactions were electrophoresed through a 1.2% agarose gel, Southern blotted to Hybond N (Amersham International plc) and hybridised with the same primers used in the PCR reaction, 5' end-labelled with ($\gamma^{32}$P)ATP. The Southern blots were washed at medium stringency.
Autoradiographic results obtained from s-q rtPCR analysis of cpr-5

Each of 20 cDNA samples, generated from total RNA isolated at two hour intervals post hatch, was amplified by PCR. Each PCR reaction was performed using primers specific for cpr-5 and the internal control gene, ama-1. The PCR reactions were electrophoresed through a 1.2% agarose gel, Southern blotted to Hybond N (Amersham International plc) and hybridised with the same primers used in the PCR reaction, 5' end-labelled with (γ³²P)ATP. The Southern blots were washed at medium stringency.
Figure 5.8

Autoradiographic results obtained from s-q rtPCR analysis of *cpr-6*

Each of 20 cDNA samples, generated from total RNA isolated at two hour intervals post hatch, was amplified by PCR. Each PCR reaction was performed using primers specific for *cpr-6* and the internal control gene, *ama-1*. The PCR reactions were electrophoresed through a 1.2% agarose gel, Southern blotted to Hybond N (Amersham International plc) and hybridised with the same primers used in the PCR reaction, 5' end-labelled with (γ^{32}P)ATP. The Southern blots were washed at medium stringency.
Figure 5.9

Autoradiographic results obtained from s-q rtPCR analysis of \textit{dpy-7}

The autoradiographic results of s-q rtPCR analysis of \textit{dpy-7} expression were obtained by Dr. I.L. Johnstone who also donated the autoradiograph shown in this figure.
Figure 5.10

A summary of the modified s-q rtPCR approach used to analyse the temporal patterns of expression of *cpr-3*, *cpr-4*, *cpr-5* and *cpr-6*

Notes: Generation of the cDNA samples:

The Embryo cDNA was generated from RNA extracted from mixed stage embryos. The L₁, L₂, L₃, L₄ and Adult cDNA samples were generated by pooling the appropriate samples from the 20 staged cDNA samples generated from total RNA isolated at two hour intervals post embryo hatch.

The samples were pooled in the following manner:

- **L₁:** Samples 1-6 (from 2-12 hours post hatch)
- **L₂:** Samples 7-9 (from 12-18 hours post hatch)
- **L₃:** Samples 10-12 (from 18-24 hours post hatch)
- **L₄:** Samples 13-15 (from 24-30 hours post hatch)
- **Adult:** Samples 16-20 (from 30-40 hours post hatch)
cDNA samples from Embryo (E), L1, L2, L3, L4 and Adult (A) stages

Three independent PCR reactions were performed on cDNA from each developmental stage. For each reaction, 23 cycles of PCR were used with primers specific for the test gene and for ama-1 (the internal control gene)

- Reaction A
  - Amplified
  - E, L1, L2, L3, L4, A

- Reaction B
  - Amplified
  - E, L1, L2, L3, L4, A

- Reaction C
  - Amplified
  - E, L1, L2, L3, L4, A

Each PCR reaction was divided into two 20μl aliquots and electrophoresed through an agarose gel

- Each gel was cut in half and each half Southern blotted to Hybond N

- Filter 1
- Filter 2
- Filter 3
- Filter 4
- Filter 5
- Filter 6

Each filter was probed simultaneously using probes specific for the target gene and ama-1.

The filters were washed at high stringency and autoradiographed

Autoradiographs were used as templates to excise the bands from the filters which corresponded to the amplification products of target cDNA and ama-1 cDNA. The radioactivity of the bands was measured by scintillation for statistical analysis.
Figure 5.11

Autoradiographic results obtained from s-q rtPCR analysis of cpr-3 using the modified protocol

A summary of the protocol is shown in Figure 5.10.

Filters 1 and 2 were obtained from independent Southern blots of the same PCR reaction (Reaction A).
Filters 3 and 4 were obtained from independent Southern blots of the same PCR reaction (Reaction B).
Filters 5 and 6 were obtained from independent Southern blots of the same PCR reaction (Reaction C).

The template from which each amplification product was derived is indicated.

The sizes of the amplification products are as follows:

Contaminating genomic DNA amplified with the cpr-3 primers: 1070bp

cDNA derived from cpr-3 transcript amplified with the cpr-3 primers: 664bp

Contaminating genomic DNA amplified with the ama-1 primers: 495bp

cDNA derived from ama-1 transcript amplified with the ama-1 primers: 340bp
Figure 5.12

Autoradiographic results obtained from s-q rtPCR analysis of *cpr-4* using the modified protocol

A summary of the protocol is shown in Figure 5.10.

Filters 1 and 2 were obtained from independent Southern blots of the same PCR reaction (Reaction A).

Filters 3 and 4 were obtained from independent Southern blots of the same PCR reaction (Reaction B).

Filters 5 and 6 were obtained from independent Southern blots of the same PCR reaction (Reaction C).

The template from which each amplification product was derived is indicated.

The sizes of the amplification products are as follows:

cDNA derived from *cpr-4* transcript amplified with the *cpr-4* primers: 418bp

Contaminating genomic DNA amplified with the *ama-1* primers: 495bp

cDNA derived from *ama-1* transcript amplified with the *ama-1* primers: 340bp

**Note:** Filters 3 and 6 were photographed in the reverse orientation.
Filter 1

ama-1 genomic DNA

Filter 2

cpr-4 transcript

ama-1 transcript

Filter 3

Filter 4

Filter 5

Filter 6

ama-1 genomic DNA

cpr-4 transcript

ama-1 transcript
Figure 5.13

Autoradiographic results obtained from s-q rtPCR analysis of *cpr-5* using the modified protocol

A summary of the protocol is shown in Figure 5.10.

Filters 1 and 2 were obtained from independent Southern blots of the same PCR reaction (Reaction A).
Filters 3 and 4 were obtained from independent Southern blots of the same PCR reaction (Reaction B).
Filters 5 and 6 were obtained from independent Southern blots of the same PCR reaction (Reaction C).

The template from which each amplification product was derived is indicated.

The sizes of the amplification products are as follows:
Contaminating genomic DNA amplified with the *cpr-5* primers: 1004bp
cDNA derived from *cpr-5* transcript amplified with the *cpr-5* primers: 770bp
Contaminating genomic DNA amplified with the *ama-1* primers: 495bp
 cDNA derived from *ama-1* transcript amplified with the *ama-1* primers: 340bp
Autoradiographic results obtained from s-q rtPCR analysis of *cpr-6* using the modified protocol

A summary of the protocol is shown in Figure 5.10.

Filters 1 and 2 were obtained from independent Southern blots of the same PCR reaction *(Reaction A)*.
Filters 3 and 4 were obtained from independent Southern blots of the same PCR reaction *(Reaction B)*.
Filters 5 and 6 were obtained from independent Southern blots of the same PCR reaction *(Reaction C)*.

The template from which each amplification product was derived is indicated.

The sizes of the amplification products are as follows:
- Contaminating genomic DNA amplified with the *cpr-6* primers: 1295bp
- cDNA derived from *cpr-6* transcript amplified with the *cpr-6* primers: 960bp
- Contaminating genomic DNA amplified with the *ama-1* primers: 495bp
- cDNA derived from *ama-1* transcript amplified with the *ama-1* primers: 340bp
Graphical representations of the relative transcript abundance ratios obtained for \(cpr-3\) (graphs A1 and A2), \(cpr-4\) (graphs B1 and B2), \(cpr-5\) (graphs C1 and C2) and \(cpr-6\) (graphs D1 and D2) using s-q rtPCR

A summary of the protocol is shown in Figure 5.10.

Each graph shows the relative transcript abundance ratio (Y-axis) for each developmental stage (X-axis). Each pair of graphs was generated from the radioactivity data obtained for \(cpr-3\), \(cpr-4\), \(cpr-5\) and \(cpr-6\), shown in Tables 5.1, 5.2, 5.3 and 5.4 respectively. Graphs A1, B1, C1 and D1 show the relative transcript abundance ratios for \(cpr-3\), \(cpr-4\), \(cpr-5\) and \(cpr-6\) respectively, obtained from filters 1, 3 and 5 for each gene*. Graphs A2, B2, C2 and D2 show the relative transcript abundance ratios for \(cpr-3\), \(cpr-4\), \(cpr-5\) and \(cpr-6\) respectively, obtained from filters 2, 4 and 6 for each gene.

Each pair of graphs shows data generated from the three independent sets of PCR reactions (A, B and C, Figure 5.10) performed for each gene.

**KEY:**

Relative transcript abundance ratios determined from PCR reaction A (filters 1 and 2, for each gene):

```
----------------------------------
```

Relative transcript abundance ratios determined from PCR reaction B (filters 3 and 4, for each gene):

```
-----------------------------
```

Relative transcript abundance ratios determined from PCR reaction C (filters 5 and 6, for each gene):

```
- - - - - -
```

* For \(cpr-6\), the radioactivity data from filters 3 and 5 were lost and therefore only the relative transcript abundance ratios obtained from filter 1 are shown.
The temporal patterns of expression of \textit{cpr-3} (graph A), \textit{cpr-4} (graph B), \textit{cpr-5} (graph C) and \textit{cpr-6} (graph D) generated by s-q rtPCR.

A summary of the protocol is shown in Figure 5.10. Each graph shows the mean and standard deviation of the relative transcript abundance ratio (Y-axis) for each developmental stage (X-axis), determined from three independent PCR reactions. The radioactivity data used to generate graphs A, B, C and D are shown in Tables 4.1, 4.2, 4.3 and 4.4 respectively. For graphs A, B and C, the mean and standard deviation of the relative transcript abundance ratio for each developmental stage was calculated from radioactivity data obtained from filters 1, 3 and 5. For graph D, the mean and standard deviation of the relative transcript abundance ratio for each developmental stage was calculated from radioactivity data obtained from filters 2, 4 and 6.
Figure 5.17

A schematic diagram showing the features of the pPD21.28 and pPD95.03 lacZ fusion vectors used to study the spatial expression patterns of cpr-3, cpr-4, cpr-5 and cpr-6

The figure shows the relevant features of the pPD21.28 vector. The pPD95.03 vector was derived from pPD21.28 and therefore only the additional features of this vector are shown.
Scheme diagram showing the schematic representation of a construct containing 12 synthetic introns. The construct consists of a Luc-5A end cassette followed by a coding sequence containing β-galactosidase coding. The diagram also includes a nuclear localization signal, synthetic intron, and polylinker. The relevant features of the construct are indicated by the additional labels on the scheme diagram.
A schematic diagram showing the approach used to obtain the fragments necessary to generate in-frame translational fusions of the 5' flank regions of cpr-3, cpr-4, cpr-5 and cpr-6. For each gene, Fragment 1 was prepared by digestion of the appropriate subclone with restriction endonucleases recognising sites ‘B’ and ‘U’ and gel purification of the appropriate fragment. Fragment 2 was generated by 15 cycles of PCR using, primers ‘X’ and ‘Y’, Pwo DNA polymerase and the appropriate subclone as template. The PCR products were gel purified, digested with restriction endonucleases recognising sites ‘A’ and ‘U’ and subcloned either into pBluescript II SK- or directly into the appropriate lacZ fusion vector. The restriction endonucleases recognising sites ‘A’, ‘B’ and ‘U’ are given in Table 5.6A. The sizes of Fragment 1 and Fragment 2 for each gene are given in Table 5.6B.

The subclones used for the preparation of Fragment 1 and Fragment 2 for each gene are as follows:

- cpr-3: CL#034/1.7/KS+
- cpr-4: ZK1055/2.8/KS+
- cpr-5: W02B2/3/KS+
- cpr-6: C25B8/3.8/KS+
The plasmid MCS gene is oriented in-frame with the lacZ reporter gene. To allow fusion of the coding region of each gene, the 3' end of Primer Y was engineered at this unique recognition site. 

Digestion of the plasmid DNA with restriction endonucleases recognizing sites A, B, and C, digests the plasmid DNA at these sites and releases the inserts and flanks of each gene using Primer X. Primer Y and Primer DNA polymerase region of the 5' flank of each gene using Primer X. Primer Y and Primer DNA polymerase PCRs amplification (15 cycles) of the sequenced each gene of the cloned 5' flank of each gene. 

Schematic representation of the cloning site. 

Digestion of the plasmid DNA with restriction endonuclease digestion of the plasmid DNA and digestion of the sequenced each gene. 

Appropriate sized fragment was generated by gel purification of the digested DNA and the restriction endonuclease.
Figure 5.19

The sequenced regions of the 5′ flanks of cpr-3, cpr-4, cpr-5 and cpr-6

The unique restriction site ‘U’ for each gene is boxed and the putative translation initiation codon is underlined.
**cpr-3 5' Flank**

-320 tccagacaatttaccaaaa -301
-300 gtgtgaccaaatgcctacattacacacacacacacgcttttgataaagctgataagttccc -241
-240 ccctctctcttttttttttttttttttttctttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttt...
Figure 5.20

The region of each of the four genes, \textit{cpr-3}, \textit{cpr-4}, \textit{cpr-5} and \textit{cpr-6} to which the 'Y' primers anneal

The putative ATG initiation codon of each gene is marked in bold. The restriction endonuclease sites 'A' are also indicated at the 5' end of each primer.
Figure 5.21

A schematic diagram summarising the approach used to insert the 5' flanks of *cpr*-3, *cpr*-4, *cpr*-5 and *cpr*-6 into the *lacZ* fusion vectors

'A', 'B' and 'U' represent the restriction endonuclease sites used to generate the *lacZ* fusion transgene constructs for each of the four genes. The actual restriction endonuclease sites used in the preparation of each construct are shown in Table 5.6A. For each gene, the 5' flank was inserted into the *lacZ* fusion vectors as two fragments (Fragment 1 and Fragment 2). An outline of how Fragment 1 and Fragment 2 were generated for each gene is given in Figure 5.18 and the sizes of these fragments are given in Table 5.6B. The figure also indicates the names, orientation and approximate positions of the oligonucleotide primers used both for sequencing the *lacZ* fusion transgene constructs and for colony PCR.
Figure 5.22

The DNA sequence of the junction between the 5’ flank of each gene and the appropriate lacZ fusion construct

Coding DNA sequence is shown in capitals while non-coding DNA sequence is shown in lowercase. The cloning site recognised by restriction endonuclease ‘A’ for each gene is marked in bold. Vertical bars indicate the open reading frame. The amino acid encoded by each codon is indicated below the appropriate codon.
A) cpr-3

Ai) Partial sequence of the 5’ flank of cpr-3 including the putative ATG translation initiation codon and partial coding sequence

-20 aaaaattgaatactaaagagaATCTCAGTGGTATTTTGCTtgagcatctctcatctcttc +40
MLKVVFL
+41 aatatttttccattatattttttctatttagAGCAATGGCCATGGCTCTTTTTTTT +100
ALFALCSAFV

Aii) Partial sequence of the 5’ flank of cpr-3 fused to the polylinker of pPD95.03

-20 aaaaattgaatactaaagagaATCTCAGTGGTATTTTGCTtgagcatctctcatctcttc +40
MLKVVFL
+41 aatatttttccattatattttttctatttagAGCAATGGCCATGGCTCTTTTTTTT +100
ALFALCSAFV

B) cpr-4

Bi) Partial sequence of the 5’ flank of cpr-4 including the putative ATG translation initiation codon and partial coding sequence

-20 tattttgtcttttcacaaaaATCGtaagttggtggtttagctgaagtttaataattgttaa +40
M
+41 tattttgtcttttcacaaaaATCGtaagttggtggtttagctgaagtttaataattgttaa +75
KYLILALAL

Bii) Partial sequence of the 5’ flank of cpr-4 fused to the polylinker of pPD95.03

-20 tattttgtcttttcacaaaaATCGtaagttggtggtttagctgaagtttaataattgttaa +40
M
+41 tattttgtcttttcacaaaaATCGtaagttggtggtttagctgaagtttaataattgttaa +40
M

C) cpr-5

Ci) Partial sequence of the 5’ flank of cpr-5 including the putative ATG translation initiation codon and partial coding sequence

-24 accttcgcagacactttctctataATCTGCGAGCTCTCTTGCTCTGCTGCTGCTG +36
MKLSAILLVAA

Cii) Partial sequence of the 5’ flank of cpr-5 fused to the polylinker of pPD21.28

-24 accttcgcagacactttctctataATCTGCGAGCTCTCTTGCTCTGCTGCTGCTG +36
MKLSAILLVAA

D) cpr-6

Di) Partial sequence of the 5’ flank of cpr-6 including the putative ATG translation initiation codon and partial coding sequence

-20 cgctgacgcctcgacgccacATGttagttttgaacttttgaaagtaatatttttaga +40
M
+41 cgctgacgcctcgacgccacATGttagttttgaacttttgaaagtaatatttttaga +40
M

Dii) Partial sequence of the 5’ flank of cpr-6 fused to the polylinker of pPD95.03

-20 cgctgacgcctcgacgccacATGttagttttgaacttttgaaagtaatatttttaga +40
M
+41 cgctgacgcctcgacgccacATGttagttttgaacttttgaaagtaatatttttaga +40
M
Expression of the *cpr-3::lacZ* transgene fusion in transgenic *C. elegans*

Strains of *C. elegans* transgenic for the *cpr-3::lacZ* transgene fusion were fixed and stained for β-galactosidase activity with X-gal. The figure shows representative examples of the staining patterns obtained for different stages of *C. elegans* development.

A. Adult  
B. L₂  
C. L₁  
D. Embryo (threesfold stage)
Figure 5.24

Expression of the cpr-4::lacZ transgene fusion in transgenic C.elegans

Strains of C.elegans transgenic for the cpr-4::lacZ transgene fusion were fixed and stained for β-galactosidase activity with X-gal. The figure shows representative examples of the staining patterns obtained for different stages of C.elegans development.

A. **Adult**

B. **L₄**

C. **L₂**

D. **Embryo (threefold stage)**
Expression of the *cpr-5::lacZ* transgene fusion in transgenic *C. elegans*

Strains of *C. elegans* transgenic for the *cpr-5::lacZ* transgene fusion were fixed and stained for β-galactosidase activity with X-gal. The figure shows representative examples of the staining patterns obtained for different stages of *C. elegans* development.

A. Adult  
B. L₃  
C. L₂  
D. Embryo (threefold stage)
Expression of the cpr-6::lacZ transgene fusion in transgenic C.elegans

Strains of C.elegans transgenic for the cpr-6::lacZ transgene fusion were fixed and stained for β-galactosidase activity with X-gal. The figure shows representative examples of the staining patterns obtained for different stages of C.elegans development.

A. Adult
B. Two L₄ larvae
C. L₃
D. L₂
Schematic diagram of part of the alimentary tract of the early L₁.

The figure shows the 20 intestinal cells of the gut and their arrangement into nine groups (int1-int9). The ring of four anterior most intestinal cells (int1) are attached to the pharyngeal-intestinal valve. The remaining 16 intestinal cells are arranged as eight pairs (int2-int9). During L₁ lethargus most of the intestinal cells undergo nuclear divisions to generate binucleated cells. However, the nuclei of the six anterior most intestinal cells (int1 and int2) do not normally divide and any of the nuclei of the four posterior most intestinal cells (int8 and int9) may also fail to divide. Thus, the intestinal cells of L₂ larvae (and all subsequent developmental stages) normally possess 30-34 nuclei. The nuclei of all 20 intestinal cells undergo a single endoreduplication event at each moult resulting in C numbers of 32 in the adult.

The figure was adapted from Sulston et al (1983).
The nuclei of the four most posterior intestinal

cells are one of which may fail to divide.

The nuclei of the six most anterior intestinal

cells do not normally divide.

Rectal valve
The cells of the rectum
The endothelial

Intestinal valve
The 20 cells of the intestine
Pharyngeal
The cells of the intestine
Chapter 6
Chapter 6
Concluding Discussion

6.1. Summary of results

The work described in this thesis indicates that *C. elegans* possesses a cathepsin B-like multigene family comprising at least five members, the four cathepsin B-like genes *cpr-3, cpr-4, cpr-5* and *cpr-6* reported here and the previously isolated gene *cpr-1* (Ray and McKerrow, 1992). DNA sequencing of the genomic, cDNA and 5' RACE clones revealed that *cpr-3, cpr-4, cpr-5* and *cpr-6* all possess diverged gene architectures, suggesting that they have arisen from ancient gene duplication events.

Some genes in *C. elegans* appear to be transcribed polycistronically in clusters, similar to the situation observed for bacterial operons (Spieth et al., 1993). Mature, monocistronic messages are generated from these polycistronic pre-mRNAs by trans-splicing events and the spliced leader SL2 is specific for trans-splicing to downstream genes in these operons (Spieth et al., 1993). A survey of DNA sequence data generated by the *C. elegans* genome project identified 31 potential operons (Zorio et al., 1994). The authors analysed 2 Mb of DNA sequence for the presence of trans-splice sites and for gene clusters. This analysis suggested that over 70% of the 266 predicted genes examined were trans-spliced and that 26% of these genes were in operons. The authors analysed seven of these potential operons for trans-splicing and found that all fitted the expectations for polycistronic transcription. RT-PCR was used to determine whether the predicted genes were transcribed and the mRNAs transspliced. These experiments revealed that the first gene of each of the putative operons tested was either trans-spliced to SL1 or not trans-spliced at all. In contrast, downstream genes were trans-spliced exclusively to SL2 or to a mixture of SL2 and SL1.

The 5' RACE approach was used to study the transcripts of *cpr-3, cpr-4, cpr-5* and *cpr-6*. The 5' RACE experiments indicated that none of these genes are trans-spliced. Thus, none of these four genes represent downstream genes in an operon and therefore, none are transcribed from the same operon. The 5' RACE experiments indicated that transcription initiation occurs within a small region upstream of and proximal to the putative ATG translation initiation codon of each gene, resulting in the
mature transcripts of these genes possessing short 5' UTRs. These results also suggest that each of the four genes is under the control of a single promoter region proximal to the 5' end of the respective gene.

Northern analysis detected a single size species of mature transcript for each gene. Each species of mature transcript was of the expected size for its respective gene. For each of the four genes, no evidence for alternative splicing was obtained within the limits of resolution and detection of Northern analysis. The results obtained from the 5' RACE experiments and from Northern analysis are in agreement with those obtained for cpr-1 (Ray and McKerrow, 1992). Northern analysis identified a single size species of transcript for cpr-1 and primer extension analysis indicated that the mature cpr-1 transcript possesses a short 5' UTR which is not trans-spliced.

The predicted amino acid sequences encoded by cpr-3, cpr-4, cpr-5 and cpr-6 display all the characteristics expected of a cathepsin B-like enzyme. In addition, phylogenetic analysis of the predicted amino acid sequences encoded by cpr-3, cpr-4, cpr-5 and cpr-6 clusters these sequences with known cathepsin B enzymes and away from similar but distinct cysteine proteases such as cathepsin L and cathepsin H. Together, these data support the assignment of cpr-3, cpr-4, cpr-5 and cpr-6 as cathepsin B-like genes.

Though the putative enzymes encoded by cpr-1, cpr-3, cpr-4, cpr-5 and cpr-6 possess features characteristic of the cathepsin B enzyme, most of these predicted proteins are highly diverged from one another. This is demonstrated by phylogenetic analysis which organises these five predicted proteins into three groups and indicates that these three groups are almost as diverged from one another as each is from the vertebrate cathepsin B enzymes. Indeed, of the five putative enzymes encoded by cpr-1, cpr-3, cpr-4, cpr-5 and cpr-6, only those encoded by cpr-4 and cpr-5 show substantially more homology to one another than to human cathepsin B. Furthermore, comparison of the predicted amino acid sequences encoded by these five genes to human cathepsin B, reveals a number of variations at amino acid residues shown to affect the specificity and activity of recombinant rat cathepsin B (Hasnian et al., 1992). These results suggest that most of the putative enzymes encoded by cpr-1, cpr-3, cpr-4, cpr-5 and cpr-6 are highly diverged from one another and may have enzyme activities and substrate specificities both distinct from human cathepsin B and distinct from each other.
The work also strongly suggests that cpr-3, cpr-4, cpr-5 and cpr-6 are all exclusively expressed in the intestinal cells of *C.elegans*. This is in agreement with *in situ* hybridisation experiments (Ray and McKerrow, 1992) and lacZ transgene fusion experiments (C.Britton, pers. comm.) conducted with cpr-1 which suggest that expression of this gene is also confined to the intestinal cells of *C.elegans*. Though cpr-3, cpr-4, cpr-5 and cpr-6 are all expressed in the intestine, they have different patterns of temporal expression (Larminie and Johnstone, 1996). Thus, the results suggest that expression of cpr-3, cpr-4 and cpr-5 is induced at the late stages of embryogenesis but that these three genes show different temporal expression patterns from one another in the subsequent larval and adult stages of development. cpr-3 relative transcript abundance remains fairly constant post hatch, suggesting that this gene may be constitutively expressed. cpr-4 relative transcript abundance is elevated during the larval stages with peaks in L1 and L4, but decline significantly in the adult by comparison to cpr-3, suggesting that cpr-4 expression is upregulated during the larval stages of *C.elegans* development. cpr-5 relative transcript abundance increases dramatically post hatch, reaches a peak at the L2 stage, remains high in the two subsequent larval stages and declines slightly in the adult. In contrast to cpr-3, cpr-4 and cpr-5, the relative transcript abundance of cpr-6 does not increase dramatically until much later in development, around the L3-L4 stages, and continues to rise through the early adult stages.

The results suggest that cpr-3, cpr-4, cpr-5 and cpr-6 are all expressed exclusively in the intestinal cells of *C.elegans* and possess distinct, but overlapping temporal patterns of expression. The predicted amino acid sequences of the four genes indicate that they encode cathepsin B-like enzymes which may have different activities and substrate specificities. Together, these data suggest a requirement for different cathepsin B-like enzymes, or combinations of such enzymes, to process or degrade different substrates, or combinations of substrates, in the intestine during different stages of *C.elegans* development. Since the gut of *C.elegans* performs an important role in digestion and is also thought to be the principle storage organ of the worm, the four cathepsin B-like enzymes may be involved in digestion and/or the degradation/processing of stored proteins. Thus, some or all of these enzymes may be directly involved in digestion by degrading proteins in the gut lumen. The different activities and substrate
specificities proposed for these enzymes may be important in this respect since the presence of more than one cathepsin B-like enzyme in the gut lumen would increase the spectrum of protein substrates which can be degraded. Alternatively, some or all of the enzymes may perform indirect, intracellular roles in digestion, by processing other enzymes that are required for digestion in the gut lumen. cpr-3, cpr-4 and cpr-5 encode the most likely candidates for digestive enzymes since these genes are all expressed throughout the larval and adult stages of C. elegans development (the stages when this nematode species feeds).

Since the intestine of C. elegans is also thought to represent the principle storage organ of the worm, some or all of these enzymes may be involved in the degradation or processing of stored proteins. In this respect, the distinct but overlapping temporal expression patterns of cpr-3, cpr-4, cpr-5 and cpr-6 may be significant since different combinations of cathepsin B-like enzymes may be required for the processing/degradation of different stored proteins at different stages during the life cycle of C. elegans. The putative cathepsin B-like enzyme encoded by cpr-6 may be a more likely candidate for such a role because this gene is not expressed at significant levels in all the feeding stages of C. elegans development and is therefore less likely to have an essential role in digestion. However, since roles in digestion and roles in protein processing are not mutually exclusive, some or all of these enzymes may perform dual functions. For example, one cathepsin B-like enzyme may process enzymes required for digestion but also process other stored proteins. Again, the different temporal patterns of expression observed for the four genes may be significant since different combinations of enzymes with different activities and substrate specificities may be able to effect different but related functions during C. elegans development. Thus one combination of enzymes may effect a digestive role while another combination may effect a role in the degradation/processing of stored proteins.

The identification of five diverged cathepsin B-like genes that are all expressed in the gut and that have distinct but overlapping temporal patterns of expression suggests that several different cathepsin B-like enzymes may be active within the same cell at any one time during C. elegans post-embryonic development. The protein substrates hydrolysed by these enzymes and the nature of the hydrolysis (i.e. complete degradation or limited proteolysis of a substrate) may be dependent on the combination of active
enzymes present. Since at least five cathepsin B-like genes may be expressed in the same intestinal cells, a large number of potential enzyme combinations are possible and many of these combinations may be functionally significant. However the potential number of enzyme combinations may be significantly reduced when one considers where the final active enzymes are located. Thus, cathepsin B-like enzymes which are secreted into the gut lumen are unlikely to have overlapping functions with those that reside within the intestinal cells. It is also possible that not all intracellular cathepsin B-like enzymes will have overlapping functions since they may not be targeted to the same organelles. This is perhaps less likely because cathepsin B-like enzymes normally have acidic pH optima and are therefore associated with the lysosomal/endosomal compartment (or specialised organelles which are associated with this compartment, such as secretory vesicles). However, since we have not biochemically characterised these enzymes nor studied their subcellular distribution, such a possibility cannot be discounted.

The presence of a cathepsin B-like multigene family in *C. elegans* whose members exhibit spatially restricted patterns of expression is similar to the situation observed in parasitic nematode and trematode species, where cathepsin B-like multigene families have been identified (Chapter 1, Section 1.9) and evidence of tissue-restricted patterns of expression obtained (Chapter 1, Section 1.8). By comparison, vertebrates appear to possess only a single cathepsin B gene that is expressed in many, if not all, tissues of the body. Furthermore, unlike the cathepsin B-like genes of *C. elegans*, the vertebrate cathepsin B gene appears to be under the control of a complex regulatory mechanism, with evidence of differential expression from more than one promoter and of alternative splicing of the resulting transcripts. Together, these observations suggest that the cathepsin B-like enzymes of both free living and parasitic nematode and trematode species may be required for different biological functions than the cathepsin B enzyme of vertebrates.

The identification of cathepsin B-like multigene families in both a free-living nematode species and in parasitic nematode and trematode species does not preclude such enzymes performing parasite-specific roles since the biological roles of such enzymes in free-living nematode species may have been adapted for a parasitic lifestyle. However, it does indicate that the amplification of cathepsin B-like multigene families is not unique to parasitic nematode and trematode species and therefore indicates that these
enzymes may be required for general nematode and trematode biology, and not just for parasite-specific processes.

6.2. Future work

6.2.1. Determining the distribution of the cathepsin B-like enzymes

In order to determine the distribution of cathepsin B-like enzymes in *C. elegans*, it would be necessary to generate antisera specific for each of the enzymes. Specific antisera may be obtained by one of two routes. The first route would be to use synthetic peptides to immunise rabbits. There are several advantages to using such an approach to obtain antisera specific for each cathepsin B-like enzyme. The extensive knowledge of the human cathepsin B enzyme tertiary structure could be used to predict the tertiary structure of each cathepsin B-like enzyme. This would allow more accurate prediction of synthetic peptides likely to be immunogenic for each of the cathepsin B-like enzymes. Such an approach would also allow the rational selection of synthetic peptides from regions which are not conserved between each of the enzymes and so increase the probability of obtaining specific antisera for each cathepsin B-like enzyme. The second approach would be to express the cDNA for each gene in *E. coli* and use the expressed proteins to immunise rabbits. Such an approach has the disadvantage that it is likely to produce antisera for each enzyme which will cross react with other members of the cathepsin B-like multigene family. This may be overcome by expressing only portions of each cDNA which encode regions of variability between the four enzymes.

Were specific antisera to be obtained, there are a number of studies which could be performed. First, the antisera could be used for whole worm antibody staining. Such experiments would allow confirmation of the data obtained with the lacZ fusion transgenes and allow analysis of the distribution of the protein products of the four genes. In the latter case, this could determine which, if any, of the four genes may encode cathepsin B-like enzymes with a direct or indirect role in digestion by demonstrating their presence in the lumen of the intestine. Such studies would not allow cathepsin B-like enzymes with a direct role in digestion to be distinguished from those with indirect roles since both types of role are likely to result in the presence of such
enzymes in the lumen. Cathepsin B-like enzymes with a direct role in digestion will be localised in the intestinal lumen for obvious reasons. However, cathepsin B-like enzymes that perform indirect roles, by processing proteins required for digestion, may perform such roles in secretory granules and may therefore be released into the lumen with the processed proteins.

Since the results suggest that cpr-3, cpr-4, cpr-5 and cpr-6 are all expressed in the intestine of C. elegans, it would be informative to use specific antisera for immunogold electron microscopy. This approach would enable the subcellular localisation of each enzyme to be studied in detail. Furthermore, by using different sizes of gold particles, it would be possible to determine which, if any, of these enzymes are co-localised to the same organelles. The results from such studies could thereby identify cathepsin B-like enzymes which may perform overlapping functions within the same subcellular compartment.

6.2.2. Biochemical characterisation of the cathepsin B-like enzymes

Specific antisera for each of the four cathepsin B-like enzymes may also allow purification of the enzymes by antibody affinity chromatography. The free living existence of C. elegans means there is no real limit to the amount of material which can be obtained and therefore such an approach may allow the purification of sufficient quantities of each enzyme for extensive biochemical characterisation. Biochemical characterisation of the proteins encoded by cpr-3, cpr-4, cpr-5 and cpr-6 will form an important part of any future work. The proteins encoded by these genes have been termed cathepsin B-like by virtue of their homology to human cathepsin B and because they possess many of the amino acid sequence characteristics expected of cathepsin B-like enzymes. However, biochemical characterisation is essential in order to prove that these encoded proteins have cathepsin B-like activities. First, the activities and substrate preferences of the four enzymes would be studied using both protein substrates and fluorogenic synthetic peptide substrates. Protein substrates would be used to determine whether these enzymes do genuinely have different activities and substrate specificities, as suggested by their predicted amino acid sequences. Most notably, this approach would determine which of the enzymes possess the peptidyl-dipeptidase activity
associated with vertebrate cathepsin B. Banks of fluorogenic synthetic peptide substrates possessing different amino acid combinations would be used to analyse the amino acid substrate preferences of the four enzymes. In particular, the ability of these enzymes to cleave synthetic substrates containing two basic amino acid residues would be tested, since cleavage of such substrates represents another characteristic feature of vertebrate cathepsin B. In addition, the screening of such banks might reveal synthetic substrates which are specific for each enzyme. These specific synthetic substrates would allow studies similar to those performed with parasitic nematode species. Thus the E/S products from liquid cultures of *C.elegans* and the proteins from whole worm homogenates could be analysed using the specific synthetic substrates. This might identify which enzymes are active outwith *C.elegans* and therefore might identify which, if any, of these cathepsin B-like enzymes are likely to have a direct role in digestion. It should be noted that such studies would also require the isolation of the enzyme encoded by the *cpr-1* gene (Ray and McKeerrow, 1992) so that it could be tested to ensure that it did not possess any significant activity against the chosen synthetic substrates. Even after such studies, the results would have to be interpreted with some caution because one could not exclude the possibility of the presence of additional uncharacterised cathepsin B-like enzymes in *C.elegans* with similar synthetic substrate preferences to the enzymes being analysed.

The second area of biochemical characterisation would be to analyse the enzyme kinetics of each of the proteins encoded by *cpr-3, cpr-4, cpr-5* and *cpr-6*. Fluorogenic substrates would be used to determine the pH at which each enzyme shows optimal activity and the pH range within which each enzyme is active. Cysteine protease inhibitors would be used to demonstrate that these enzymes are indeed cysteine proteases. Small synthetic peptide inhibitors would also be used to study the rates of inactivation of each enzyme. The specificity of these inhibitors is due to the peptide portion of the inhibitor which is bound by the subsites of the active site cleft. Thus, by screening banks of inhibitors which possess different combinations of amino acid residues, it would be possible to analyse the subsite preferences of each enzyme. Such a study would complement the results obtained from screening banks of fluorogenic synthetic peptide substrates.
Expression of recombinant rat cathepsin B in yeast results in the generation of active enzyme (Lee et al., 1990) and this enzyme has been shown to be essentially identical to rat liver cathepsin B by detailed biochemical analysis (Hasnain et al., 1992). Accordingly, it may also be possible to obtain active recombinant cathepsin B-like enzymes by expressing the cDNA of each of the four genes in yeast. This system has the advantage that it does not require specific antisera and thus would allow biochemical characterisation of the cathepsin B-like enzymes encoded by cpr-3, cpr-4, cpr-5 and cpr-6 regardless of whether or not specific antisera against these enzymes could be obtained. However, since no biochemical comparisons between the recombinant enzymes and the endogenous enzymes could be made, the results from such studies would have to be treated with a certain degree of caution.

6.2.3. The effect of starving on expression of cpr-3, cpr-4, cpr-5 and cpr-6

The gut-specific expression of cpr-3, cpr-4, cpr-5 and cpr-6 suggests that these genes may be involved in digestion. Thus expression of at least some of these genes may only be induced in the presence of a food source. This aspect of gene regulation could be studied by repeating the s-q rtPCR experiments using RNA isolated from embryos and L1 larvae cultured in the presence and absence of an E.coli food source. The identification of cathepsin B-like genes which are only expressed in the presence of a food source would provide strong evidence for a role in digestion for such genes. By the same token, it would also be useful to use s-q rtPCR to determine whether or not cpr-3, cpr-4, cpr-5 and cpr-6 are expressed in the non-feeding dauer larvae. Such studies would also demonstrate the presence of complex regulatory mechanisms which ensure that expression of such genes only occurs in the appropriate conditions.

6.2.4. Obtaining C.elegans mutant strains for cpr-3, cpr-4, cpr-5 and cpr-6

The final area of future research highlights the experimental tractability of C.elegans. Strains of C.elegans with mutations in cpr-3, cpr-4, cpr-5 or cpr-6 may be obtained by one of two routes. First, these genes may already be defined by genetic loci with known genetic map positions, determined by classical genetic analysis.
Accordingly, the alignment of the physical and genetic maps of *C. elegans* may allow the isolation of mutant strains for some or all of these genes by plasmid rescue with the cloned wild-type genes (Chapter 3, Section 3.3). If such an approach failed, mutant strains could be obtained using a reverse genetic approach. This would be achieved using the site-selected transposon mutagenesis technique outlined in Chapter 1, Section 1.3. The mutants for each of the four genes obtained by these approaches may possess phenotypes which will help to determine the biological roles for each of the four genes. Alternatively, such mutants may not possess a visible mutant phenotype. This may be a result of functional redundancy within the cathepsin B-like multigene family. Such redundancy could be analysed by constructing crosses to generate strains with mutations in two or more of the cathepsin B-like genes. Such studies could not only identify genes with overlapping functions but also gene combinations essential for nematode survival. This information may be particularly useful for the effective design of chemotherapeutic agents against parasitic nematode species.
References
References


Appendix A

The 5' RACE experiments identified a potential splice acceptor site (TTTTTCAG) in the 5' UTR of the cpr-3 mature transcript. This suggests that the mature transcript of cpr-3 may be trans-spliced to one of the two splice leaders, SL1 or SL2. Dr. I.Hope (pers. comm.) has reported problems amplifying SL1 or SL2 trans-spliced messages using a 5' RACE approach. Therefore, it is possible that the 5' RACE products generated for cpr-3 may be derived from immature mRNA transcripts which are subsequently trans-spliced. This possibility should be addressed by PCR amplification of cpr-3, cpr-4, cpr-5 and cpr-6 from a C.elegans cDNA library using the SL1 and SL2 primers and gene specific primers for each of the four genes.